Multiagent Deontic Logic and its Challenges from a Normative Systems Perspective

Gabriella Pigozzi
Université Paris-Dauphine, PSL Research University, CNRS, LAMSADE, 75016 Paris, France
gabriella.pigozzi@dauphine.fr

Leendert van der Torre
University of Luxembourg, Maison du Nombre, 6, Avenue de la Fonte, L-4364 Esch-sur-Alzette
leon.vandertorre@uni.lu

Abstract

This article gives an overview of several challenges studied in deontic logic, with an emphasis on challenges involving agents. We start with traditional modal deontic logic using preferences to address the challenge of contrary-to-duty reasoning, and STIT theory addressing the challenges of non-deterministic actions, moral luck and procrastination. Then we turn to alternative norm-based deontic logics detaching obligations from norms to address the challenge of Jørgensen's dilemma, including the question how to derive obligations from a normative system when agents cannot assume that other agents comply with their norms. We discuss also some traditional challenges from the viewpoint of normative systems: when a set of norms may be termed 'coherent', how to deal with normative conflicts, how to combine normative systems and traditional deontic logic, how various kinds of permission can be accommodated, how meaning postulates and counts-as conditionals can be taken into account,

The authors thank Jan Broersen and Jörg Hansen for their joint work on earlier versions of some sections of this article, and Davide Grossi and Xavier Parent for insighful and helpful comments on a preliminary version of this article. The contribution of G. Pigozzi was supported by the Deutsche Forschungsgemeinschaft (DFG) and the Czech Science Foundation (GACR) as part of the joint project From Shared Evidence to Group Attitudes (RO 4548/6-1). This work is supported by the European Union's Horizon 2020 research and innovation programme under the Marie Curie grant agreement No: 690974 (Mining and Reasoning with Legal Texts, MIREL).

how sets of norms may be revised and merged, and how normative systems can be combined with game theory. The normative systems perspective means that norms, not ideality or preference, should take the central position in deontic semantics, and that a semantics that represents norms explicitly provides a helpful tool for analysing, clarifying and solving the problems of deontic logic. We focus on the challenges rather than trying to give full coverage of related work, for which we refer to the handbook of deontic logic and normative systems.[1]

Introduction

Deontic logic [116, 34] is the field of logic that is concerned with normative concepts such as obligation, permission, and prohibition. Alternatively, a deontic logic is a formal system capturing the essential logical features of these concepts. Typically, a deontic logic uses Op to mean that it is obligatory that p, (or it ought to be the case that p), and Pp to mean that it is permitted, or permissible, that p. The term 'deontic' is derived from the ancient Greek *déon*, meaning that "which is binding or proper".

Deontic logic can be used for reasoning about normative multiagent systems, i.e. about multiagent systems with normative systems in which agents can decide whether to follow the explicitly represented norms, and the normative systems specify how and to which extent agents can modify the norms [16, 6]. Normative multiagent systems need to combine normative reasoning with agent interaction, and thus raise the challenge to relate the logic of normative systems to game theory [109].

Traditional (or "standard") deontic logic is a normal propositional modal logic of type KD, which means that it extends the propositional tautologies with the axioms $K : O(p \to q) \to (Op \to Oq)$ and $D : \neg(Op \wedge O\neg p)$, and it is closed under the inference rules *modus ponens* $p, p \to q/q$ and *generalization* or *necessitation* p/Op. Prohibition and permission are defined by $Fp = O\neg p$ and $Pp = \neg O\neg p$. Traditional deontic logic is an unusually simple and elegant theory. An advantage of its modal-logical setting is that it can easily be extended with other modalities such as epistemic or temporal operators and modal accounts of action. In this article we illustrate the combination of deontic logic with a modal logic of action, called STIT logic [58].

Not surprisingly for such a highly simplified theory, there are many features of actual normative reasoning that traditional deontic logic does not capture. Noto-

[1] Sections 2-4 are based on a review of Horty's book on obligation and agency [23], Section 1 and Sections 5-14 are based on a technical report of a Dagstuhl seminar [52], and Section 15 is based on an article of the second author of this paper [109].

rious are the so-called 'paradoxes of deontic logic', which are usually dismissed as consequences of the simplifications of traditional deontic logic. For example, Ross's paradox [99] is the counterintuitive derivation of "you ought to mail or burn the letter" from "you ought to mail the letter." It is typically viewed as a side effect of the interpretation of 'or' in natural language.

In this article we discuss also an example of norm based semantics, called input/output logic, to discuss challenges related to norms and detachment. Maybe the most striking feature of the abstract character of traditional deontic logic is that it does not explicitly represent the norms of the system, only the obligations and permissions which can be detached from the norms in a given context. This is an obvious limitation when using deontic logic to reason about normative multiagent systems, in which norms are represented explicitly.

In this article we consider the following fifteen challenges for multiagent deontic logic. The list of challenges is by no means final. Other problems may be considered equally important, such as how a hierarchy of norms (or of the norm-giving authorities) is to be respected, how general abstract norms relate to individual concrete obligations, how norms can be interpreted, or how various kinds of imperatives can be distinguished. We do not consider deontic logics for specification and verification of multiagent systems [20, 1], but we focus on normative reasoning within multiagent systems. The three central concepts in these challenges are preference, agency, and norms. Regarding agency, we consider individual agent action as well as agent interaction in games.

1.	Contrary-to-duty reasoning, preference and violation	preference
2.	Non-deterministic actions: ought-to-do vs ought-to-be	agency
3.	Moral luck and the driving example	agency
4.	Procrastination: actualism vs possibilism	agency
5.	Jørgensen's dilemma and the problem of detachment	norms
6.	Multiagent detachment	norms
7.	Coherence of a normative system	norms
8.	Normative conflicts and dilemmas	preference & norms
9.	Descriptive dyadic obligations and norms	preference & norms
10.	Permissive norms	preference & norms
11.	Meaning postulates and intermediate concepts	norms
12.	Constitutive norms	norms
13.	Revision of a normative system	norms
14.	Merging normative systems	norms
15.	Games, norms and obligations	norms & agency

To discuss these challenges, we repeat the basic definitions of so-called standard deontic logic, dyadic standard deontic logic, deontic STIT logic, and input/output logic. The article thus contains several definitions, but these are not put to work in any theorems or propositions, for which we refer to the handbook of deontic logic and normative systems [34]. The point of introducing formal definitions in this article is just to have a reference for the interested reader. Likewise, the interested reader should consult the handbook of deontic logic and normative systems for a more comprehensive description of the work done on each challenge, as in this article we can mention only a few references for each challenge.

1 Contrary-to-duty reasoning, preference and violation

In this section we discuss how the challenge of the contrary-to-duty paradoxes leads to traditional modal deontic logic introduced at the end of the sixties, based on dyadic operators and preference based semantics. Moreover, we contrast this use of preference in deontic logic with the use of preference in decision theory.

1.1 Chisholm's paradox

Suppose we are given a code of conditional norms, that we are presented with a condition (input) that is unalterably true, and asked what obligations (output) it gives rise to. It may happen that the condition is something that should not have been true in the first place. But that is now water under the bridge: we have to "make the best out of the sad circumstances" as B. Hansson [53] put it. We therefore abstract from the deontic status of the condition, and focus on the obligations that are consistent with its presence. How to determine this in general terms, and if possible in formal ones, is the well-known problem of contrary-to-duty conditions as exemplified by the notorious contrary-to-duty paradoxes. Chisholm's paradox [28] consists of the following four sentences:

(1) It ought to be that a certain man go to the assistance of his neighbours.
(2) It ought to be that if he does go, he tell them he is coming.
(3) If he does not go then he ought not to tell them he is coming.
(4) He does not go.

Furthermore, intuitively, the sentences derive the following sentence (5):

(5) He ought not to tell them he is coming.

Chisholm's paradox is a contrary-to-duty paradox, since it contains both a primary obligation to go, and a secondary obligation not to tell if the agent does not go. Traditionally, the paradox was approached by trying to formalise each of the

sentences in an appropriate language of deontic logic. However, in traditional (or "standard") deontic logic, i.e. the normal propositional modal logic of type KD, it turned out that either the set of formulas is inconsistent, or one formula is a logical consequence of another formula. Yet intuitively the natural-language expressions that make up the paradox are consistent and independent from each other: this is why it is called a paradox. The problem is thus:

Challenge 1. *How do we reason with contrary-to-duty obligations which are in force only in case of norm violations?*

There are various kinds of scenarios which are similar to Chisholm's scenario. For example, there is a key difference between contrary-to-duties proper, and reparatory obligations, because the latter cannot be atemporal [98]. Though Chisholm presented his challenge as essentially a single agent decision problem, we can as well reformulate it as a multiagent reasoning problem:

(1) It is obligatory that i sees to it that p (i should do p).
(2) It is obligatory that j sees to it that q if i does not see to it that p (j should sanction i if i does not do as told).
(3) It is obligatory that j does not see to it that q if i sees to it that p (j should not sanction i if i does as told).
(4) i does not do as told.

The logic may give us the paradoxical conclusion that j should see to it that q and he should see to it that not q. For example, van Benthem, Grossi and Liu [108] give the following example, in the formulation proposed by Åqvist [7]:

(1) It ought to be that Smith refrains from robbing Jones.
(2) Smith robs Jones.
(3) If Smith robs Jones, he ought to be punished for robbery.
(4) It ought to be that if Smith refrains from robbing Jones he is not punished for robbery.

As explained in detail in the following subsections, the development of dyadic deontic operators as well as the introduction of temporally relative deontic logic operators can be seen as a direct result of Chisholm's paradox. Since the robbing takes place before the punishment, the example can quite easily be represented once time is made explicit [110]. If you make time explicit or you direct obligations to different agents, then the paradox disappears, in a way. However, both the fact that time and agency are present may distract from the key point behind the example. Therefore also atemporal, non-agency version of the paradox allow to address to the core challenge of the issue. For example, Prakken and Sergot [98] consider the following variant of Chisholm's scenario:

(1) It ought to be that there is no dog.
(2) If there is a dog, there should be a sign.
(3) If there is no dog, there should be no sign.
(4) There is a dog.

When a new deontic logic is proposed, the traditional contrary-to-duty examples are always the first benchmark examples to be checked. It may be observed here that some researchers in deontic logic doubt that contrary-to-duties can still be considered a challenge, because due to extensive research by now we know pretty much everything about them. The deontic logic literature is full of (at least purported) solutions. In other words, these researchers doubt that deontic logic still needs more research on contrary-to-duties. Indeed, it appears to be difficult to make an original contribution to this vast literature, but new twists are still identified [96].

1.2 Monadic deontic logic

Traditional or 'standard' deontic logic, often referred to as SDL, was introduced by Von Wright [116].

1.2.1 Language

Let Φ be a set of propositional letters. The language of traditional deontic logic $\mathfrak{L}_D$ is given by the following BNF:

$$\varphi := \bot \mid p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid \bigcirc\varphi \mid \Box\varphi$$

where $p \in \Phi$. The intended reading of $\bigcirc\varphi$ is "φ is obligatory" and the intended reading of $\Box\varphi$ is "φ is necessary". Moreover we use $P\varphi$, read as "φ is permitted", as an abbreviation of $\neg \bigcirc \neg\varphi$ and $F\varphi$, "φ is forbidden", as an abbreviation of $\bigcirc\neg\varphi$. Likewise, $\vee$, $\rightarrow$ and $\leftrightarrow$ are defined in the usual way.

1.2.2 Semantics

The semantics is based on an accessibility relation that gives all the ideal alternatives of a world.

Definition 1.1. *A deontic relational model $M = (W, R, V)$ is a structure where:*

- *W is a nonempty set of worlds.*
- *R is a serial relation over W. That is, $R \subseteq W \times W$ and for all $w \in W$, there exist $v \in W$ such that Rwv.*

- *V is a valuation function that assigns a subset of W to each propositional letter p. Intuitively, $V(p)$ is the set of worlds in which p is true.*

A formula $\bigcirc\varphi$ is true at world w when φ is true in all the ideal alternatives of w.

Definition 1.2. *Given a relational model M, and a world s in M, we define the satisfaction relation $M, s \models A$ ("world s satisfies A in M") by induction on A using the clauses:*

- $M, s \vDash p$ *iff* $s \in V(p)$.
- $M, s \vDash \neg\varphi$ *iff not* $M, s \vDash \varphi$.
- $M, s \vDash (\varphi \wedge \psi)$ *iff* $M, s \vDash \varphi$ *and* $M, s \vDash \psi$.
- $M, s \vDash \bigcirc\varphi$ *iff for all* t, *if* Rst *then* $M, t \vDash \varphi$.
- $M, s \vDash \Box\varphi$ *iff for all* $t \in W$, $M, t \vDash \varphi$.

For a set Γ of formulas, we write $M, s \vDash \Gamma$ iff for all $\varphi \in \Gamma$, $M, s \vDash \varphi$. For a set Γ of formulas and a formula φ, we say that φ is a consequence of Γ (written as $\Gamma \vDash \varphi$) if for all models M and all worlds $s \in W$, if $M, s \vDash \Gamma$ then $M, s \vDash \varphi$.

1.2.3 Limitations

The following example is a variant of the scenario originally phrased by Chisholm in 1963. There is widespread agreement in the literature that, from the intuitive point of view, this set of sentences is consistent, and its members are logically independent of each other.

(A) It ought to be that Jones does not eat fast food for dinner.

(B) It ought to be that if Jones does not eat fast food for dinner, then he does not go to McDonald's.

(C) If Jones eats fast food for dinner, then he ought to go to McDonald's.

(D) Jones eats fast food for dinner.

Below are three ways to formalise this example. The first attempt is inconsistent. The second attempt is redundant due to $\bigcirc\neg f \models \bigcirc(f \to m)$. The third attempt is redundant due to $f \models \neg f \to \bigcirc\neg m$.

(A_a)	$\bigcirc\neg f$	(A_b)	$\bigcirc\neg f$	(A_c)	$\bigcirc\neg f$
(B_a)	$\bigcirc(\neg f \to \neg m)$	(B_b)	$\bigcirc(\neg f \to \neg m)$	(B_c)	$\neg f \to \bigcirc\neg m$
(C_a)	$f \to \bigcirc m$	(C_b)	$\bigcirc(f \to m)$	(C_c)	$f \to \bigcirc m$
(D_a)	f	(D_b)	f	(D_c)	f

However, it is not very hard to meet the two requirements of consistency and logical independence. The following representation is an example. It comes with apparently strong assumptions, because B_1/C_1 seem to say that my (conditional) obligations are necessary. For instance, Anderson argued that norms are contingent, because we make our rules; they are not (logical) necessities. However, we could also say that the $\Box$ is just part of the definition of a strict conditional. Also, we could represent the first obligation as $\Box \bigcirc \neg f$.

(A_1) $\bigcirc \neg f$
(B_1) $\Box(\neg f \rightarrow \bigcirc \neg m)$
(C_1) $\Box(f \rightarrow \bigcirc m)$
(D_1) $\neg f$

More seriously, a drawback of the SDL representation $A_1 - D_1$ is that it does not represent that ideally, the man does not eat fast food and does not go to McDonald's. In the ideal world, Jones goes to McDonald, yet he does not eat fast food. Moreover, there does not seem to be a similar solution for the following variant of the scenario. It is a variant of Forrester's paradox [33], also known as the gentle murderer paradox: You should not kill, but if you kill, you should do it gently.

(AB) It ought to be that Jones does not eat fast food and does not go to McDonald's.

(C) If Jones eats fast food, then he ought to go to McDonald's.

(D) Jones eats fast food for dinner.

Moreover, SDL uses a binary classification of worlds into ideal/non-ideal, whereas many situations require a trade-off between violations. The challenge is to extend the semantics of SDL in order to overcome this limitation. For example, one can add distinct modal operators for primary and secondary obligations, where a secondary obligation is a kind of reparational obligation. From $A_2 - D_2$ we can derive only $\bigcirc_1 m \wedge \bigcirc_2 \neg m$, which is perfectly consistent.

(A_2) $\bigcirc_1 \neg f$
(B_2) $\bigcirc_1(\neg f \rightarrow \neg m)$
(C_2) $f \rightarrow \bigcirc_2 m$
(D_2) f

However, it may not always be easy to distinguish primary from secondary obligations, because it may depend on the context whether an obligation is primary or secondary. For example, if we leave out **A**, then **C** would be a primary obligation instead of a secondary one. Carmo and Jones [25] therefore put as an additional requirement for a solution of the paradox that **B** and **C** are represented in the same

way (as in A_1-D_1). Also, the distinction between $\bigcirc_1$ and $\bigcirc_2$ is insufficient for extensions of the paradox that seem to need also operators like $\bigcirc_3$, $\bigcirc_4$, etc, such as the following **E** and **F**.

(E) If Jones eats fast food but does not go to McDonald's, then he should go to Quick.

(F) If Jones eats fast food but does not go to McDonald's or to Quick, then he should ...

1.2.4 SDL proof system

The proof system of traditional deontic logic Λ_D is the smallest set of formulas of $\mathfrak{L}_D$ that contains all propositional tautologies, together with the following axioms:

K $\bigcirc(\varphi \to \psi) \to (\bigcirc\varphi \to \bigcirc\psi)$

D $\bigcirc\varphi \to P\varphi$

and is closed under *modus pones*, and *generalization* (that is, if $\varphi \in \Lambda_D$, then $\bigcirc\varphi \in \Lambda_D$).

For every $\varphi \in \mathfrak{L}_\mathfrak{D}$, if $\varphi \in \Lambda_D$ then we say φ is a theorem and write $\vdash \varphi$. For a set of formulas Γ and formula φ, we say φ is deducible form Γ (write $\Gamma \vdash \varphi$) if $\vdash \varphi$ or there are formulas $\psi_1, \ldots, \psi_n \in \Gamma$ such that $\vdash (\psi_1 \wedge \ldots \wedge \psi_n) \to \varphi$.

1.3 Dyadic deontic logic

Inspired by rational choice theory in the sixties, preference-based semantics for traditional deontic logic was used by, for example, Danielsson [32], Hansson [53], van Fraassen [115], Lewis [74], and Spohn [104]. The obligations of Chisholm's paradox can be represented by a preference ordering, like:

$$\neg f \wedge \neg m > \neg f \wedge m > f \wedge m > f \wedge \neg m$$

Extensions like **E** and **F** can be incorporated by further refining the preference relation. The language is extended with dyadic operators $\bigcirc(p|q)$, which is true iff the preferred q worlds satisfy p. The class of logics is called Dyadic 'Standard' Deontic Logic or DSDL. The notation is inspired by the representation of conditional probability.

1.3.1 Language

Given a set Φ of propositional letters. The language of DSDL $\mathfrak{L}_D$ is given by the following BNF:

$$\varphi := \bot \mid p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid \Box\varphi \mid \bigcirc(\varphi|\varphi)$$

The intended reading of $\Box\varphi$ is "necessarily φ", $\bigcirc(\varphi|\psi)$ is "It ought to be φ, given ψ". Moreover we use $P(\varphi|\psi)$, read as "φ is permitted, given ψ", as an abbreviation of $\neg \bigcirc (\neg\varphi|\psi)$, and $\Diamond\varphi$, read as "possibly φ", as an abbreviation of $\neg\Box\neg\varphi$.

Unconditional obligations are defined in terms of the conditional ones by $\bigcirc p = \bigcirc(p|\top)$, where $\top$ stands for any tautology.

1.3.2 Semantics

The semantics is based on an accessibility relation that gives all better alternatives of a world.

Definition 1.3. *A preference model $M = (W, \geq, V)$ is a structure where:*

- *W is a nonempty set of worlds.*
- *$\geq$ is a reflexive, transitive relation over W satisfying the following limitedness requirement: if $||\varphi|| \neq \emptyset$ then $\{x \in ||\varphi|| : (\forall y \in ||\varphi||)x \geq y\} \neq \emptyset$. Here $||\varphi|| = \{x \in W : M, x \vDash \varphi\}$.*
- *V is a standard propositional valuation such that for every propositional letter p, $V(p) \subseteq W$.*

Definition 1.4. *Formulas of $\mathfrak{L}_D$ are interpreted in preference models.*

- *$M, s \vDash p$ iff $s \in V(p)$.*
- *$M, s \vDash \neg\varphi$ iff not $M, s \vDash \varphi$.*
- *$M, s \vDash (\varphi \wedge \psi)$ iff $M, s \vDash \varphi$ and $M, s \vDash \psi$.*
- *$M, s \vDash \Box\varphi$ iff $\forall t \in W$, $M, t \vDash \varphi$.*
- *$M, s \vDash \bigcirc(\psi|\varphi)$ iff $\forall t(((M, t \vDash \varphi) \& \forall u(M, u \vDash \varphi) \Rightarrow t \geq u) \Rightarrow M, t \vDash \psi)$.*

Intuitively, $\bigcirc(\psi|\varphi)$ holds whenever the best φ-worlds are ψ-worlds.

The Chisholm's scenario can be formalised in DSDL as follows:

$(A_3)\ \bigcirc \neg f$

$(B_3)\ \bigcirc (\neg m|\neg f)$

$(C_3) \bigcirc (m|f)$

$(D_3) f$

A challenge of both the multiple obligation solution using $\bigcirc_1$, $\bigcirc_2$, ... and the preference based semantics is to combine preference orderings, for example combining the Chisholm preferences with preferences originating from the Good Samaritan paradox:

(AB') A man should not be robbed.

(C') If he is robbed, he should be helped.

(D') A man is robbed.

$$\neg r \wedge \neg h > r \wedge h > r \wedge \neg h$$

The main drawback of DSDL is that in a monotonic setting, we cannot detach the obligation $\bigcirc m$ from the four sentences. In fact, the preference based solution represents **A**, **B** and **C**, but has little to say about **D**. So the dyadic representation $A_3 - D_3$ highlights the dilemma between factual detachment (FD) and deontic detachment (DD). We cannot have both FD and DD, as we derive a dilemma $\bigcirc \neg m \wedge \bigcirc m$.

$$\frac{\bigcirc(m|f), f}{\bigcirc m} FD \qquad \frac{\bigcirc(\neg m|\neg f), \bigcirc \neg f}{\bigcirc \neg m} DD$$

1.3.3 DSDL proof system

The proof system of traditional deontic logic Λ_D, also referred as Aqvist's system G, is the smallest set of formulas of $\mathfrak{L}_D$ that contains all propositional tautologies, the following axioms. The names of the labels are taken from Parent [93]:

S5	S5-schemata for $\Box$
COK	$\bigcirc(B \to C|A) \to (\bigcirc(B|A) \to \bigcirc(C|A))$
Abs	$\bigcirc(B|A) \to \Box \bigcirc (B|A)$
CON	$\Box B \to \bigcirc(B|A)$
Ext	$\Box(A \leftrightarrow B) \to (\bigcirc(C|A) \leftrightarrow \bigcirc(C|B))$
Id	$\bigcirc(A|A)$
C	$\bigcirc(C|(A \wedge B)) \to \bigcirc((B \to C)|A)$
D *	$\Diamond A \to (\bigcirc(B|A) \to P(B|A))$

S $(P(B|A) \wedge \bigcirc((B \rightarrow C)|A)) \rightarrow \bigcirc(C|(A \wedge B))$

and is closed under *modus ponens*, and *generalization* (that is, if $\varphi \in \Lambda_D$, then $\Box\varphi \in \Lambda_D$).

1.3.4 The use of preferences in decision theory

Arrow's condition of rational choice theory says that if C are the best alternatives of A, and $B \cap C$ is nonempty, then $B \cap C$ are the best alternatives of $A \cap B$. This principle is reflected by the S axiom of DSDL:

$$(P(B|A) \wedge \bigcirc((B \rightarrow C)|A)) \rightarrow \bigcirc(C|(A \wedge B))$$

Moreover, we may represent a preference or comparative operator $\succ$ in the language, and define the dyadic operator in terms of the preference logic:

$$O(\psi \mid \phi) =_{def} (\phi \wedge \psi) \succ (\phi \wedge \neg\psi)$$

One may wonder whether the parallel between deontic reasoning and rational choice can be extended to utility theory, decision theory, game theory, planning, and so on. First, consider a typical example from Prakken and Sergot's Cottage Regulations [98]: there should be no fence, if there is a fence there should be a white fence, if there is a non-white fence, it should be black, if there is a fence which is neither white nor black, then .... This part of the cottage regulations is related to Forrester's paradox [33]. However, note the following difference between Forrester's paradox and the cottage regulations. Once you kill someone, it can no longer be undone, whereas if you build a fence, you can still remove it. The associated preferences of the fence example are:

$$\textit{no fence} > \textit{white fence} > \textit{black fence} > \ldots$$

If this represents a utility ordering over states, then we miss the representation of action [97]. For example, it may be preferred that the sun shines, but we do not say that the sun should shine. As a simple model of action, one might distinguish controllable from uncontrollable propositions [19], and restrict obligations to controllable propositions. Moreover, we may consider actions instead of states: we should remove the fence if there is one, we may paint the fence white, we may paint it black, etc.

$$\textit{remove} > \textit{paint white} > \textit{paint black} > \ldots$$

We may interpret this preference ordering as an ordering of expected utility of actions. Alternatively, the ordering may be generated by another decision rule,

such as maximin or minimal regret. Once we are working with a decision theoretic semantics, we may represent probabilities explicitly, or model causality. For example, let n stand for not doing homework and g for getting a good grade for a test. Then we may have the following preference order, which does not reflect that doing homework causes good grades:

$$n \wedge g > \neg n \wedge g > n \wedge \neg g > \neg n \wedge \neg g$$

1.3.5 The use of goals in planning and agent theory

We may interpret $O\phi$ or $O(\phi \mid \psi)$ as goals for ϕ, rather than obligations. This naturally leads to the distinction between maintenance and achievement goals, and to extensions of the logic with beliefs and intentions. Belief-Desire-Intention or BDI logics have been developed as formalizations of BDI theory.

BDI theory is developed in the theory of mind and has been based on folk psychology. In planning, more efficient alternatives to classical planning have been developed, for example based on hierarchical or graph planning.

The following example is a more challenging variant of Chisholm's scenario using anankastic conditionals [31], also known as hypothetical imperatives. The four sentences can be given a consistent interpretation, when the second sentence is interpreted as a classical conditional, and the third sentence is interpreted as an anankastic conditional.

(a) It ought to be that you do not smoke.

(b) If you want to smoke, then you should not buy cigarettes.

(c) If you want to smoke, then you should buy cigarettes.

(d) You want to smoke.

1.4 Defeasible Deontic Logic: detachment and constraints

Defeasible deontic logics (DDLs) use techniques developed in non-monotonic logic, such as constrained inference [60, 86]. Using these techniques, we can derive $\bigcirc m$ from only the first two sentences **A** and **B**, but not from all four sentences **A**-**D**. Consequently, the inference relation is not monotonic. For example, we may read $O(\phi|\psi)$ as follows: if the facts are exactly ψ, then ϕ is obligatory. This implies that we no longer have that $O(\phi)$ is represented by $O(\phi|\top)$.

In a similar fashion, in deontic update semantics (see van der Torre and Tan [111, 113, 112]) facts are updates that restrict the domain of the model. They make a fact 'settled' in the sense that it will never change again even after future

updates of the same sort. Van Benthem et al. [108] use dynamic logic to phrase such a dynamic approach within standard modal logic including reduction axioms and standard model theory. They rehabilitate classical modal logic as a legitimate tool to do deontic logic, and position deontic logic within the growing dynamic logic literature.

A drawback of the use of non-monotonic techniques is that we often have that violated obligations are no longer derived. This is sometimes referred to as the drowning problem. For example, in the cottage regulations, if it is no longer derived that there should be no fence once there is a fence, then how do we represent that a violation has occurred?

A second related drawback of this solution is that it does not give the cue for action that the decision maker should change his mind. For example, once there is a fence, it does not represent the obligation to remove the fence.

A third drawback of this approach is that the use of non-monotonic logic techniques like constraints should also be used to represent exceptions, and it thus raises the challenge how to distinguish violations from exceptions. This is highlighted by Prakken and Sergot's cottage regulations [98].

(A") It ought to be that there is no fence around the cottage.

(BC") If there is a fence around the cottage, then it ought to be white.

(G") If the cottage is close to a cliff, then there ought to be a fence.

(D") There is a fence around the cottage, which is close to a cliff.

We say more about defeasible deontic logic in Section 8.

1.5 Alternative approaches

Carmo and Jones [25] suggest that the representation of the facts is challenging, instead of the representation of the norms. In their approach, depending on the formalisation of the facts various obligations can be detached.

Another approach to Chisholm's paradox is to detach both obligations of the dilemma $\bigcirc\neg m \wedge \bigcirc m$, and represent them consistently using some kind of minimal deontic logic, for example using techniques from paraconsistent logic. From a practical reasoning point of view, a drawback of this approach is that a dilemma is not very useful as a moral cue for action. Moreover, intuitively it is not clear that the example presents a true dilemma. We say more about dilemmas in Section 9.

A recent representation of Chisholm's paradox [94, 95, 107] is to replace deontic detachment by so-called aggregative deontic detachment (ADD), and to derive from

A-D the obligation $\bigcirc(\neg f \wedge \neg m)$ and $\bigcirc m$, but not $\bigcirc \neg m$.

$$\frac{\bigcirc(m|f), f}{\bigcirc m} FD \qquad \frac{\bigcirc(\neg m|\neg f), \bigcirc \neg f}{\bigcirc(\neg m \wedge \neg f)} ADD$$

A possible drawback of these approaches is that we can no longer accept the principle of weakening (also known as inheritance).

$$\frac{\bigcirc(\neg m \wedge \neg f|\top)}{\bigcirc(\neg m|\top)} W$$

2 Non-deterministic actions: ought-to-do vs ought-to-be

We now turn to three specific challenges on agency and obligation, discussed in much more detail by Horty [58, 23]. His textbook is a prime reference for the use of deontic logic for multiagent systems. The central challenge Horty addresses is whether ought-to-do can be reduced to ought-to-be. A particular problem is the granularity of actions in case of non-deterministic effects, like flipping a coin or throwing a dice.

Challenge 2. *How to define obligations to perform non-deterministic actions?*

At first sight, we may define an obligation to do an action as an obligation that such an action is done, and we can thus reuse SDL or DSDL to define obligations regarding non-deterministic actions. In other words, it may seem that we can reduce ought-to-do to ought-to-be. However, as we discuss in Section 2.2, such a reduction is problematic. To explain this challenge, we first introduce a logic to express non-deterministic actions, so-called See-To-It-That or STIT logic.

2.1 Horty's STIT logic

We give a very brief overview of the main concepts of Horty's STIT logic. For more details and motivation we refer to Horty's textbook on obligation and agency [58]. As illustrated in Figure 1, a STIT model is a tree where each moment is a partitioning of traces or histories, where the partitioning $Choice^m_\alpha$ represents the choices of the agent at that moment. Each alternative of the choice is called an action K^m_1, K^m_2, etc. With each history a utility value is associated, and the higher the utility value, the better the history.

Formulas are evaluated with respect to moment-history pairs. Some typical formulas of Horty's utilitarian STIT-formalism are A, FA, $[\alpha\ cstit : A]$, and $\bigcirc A$

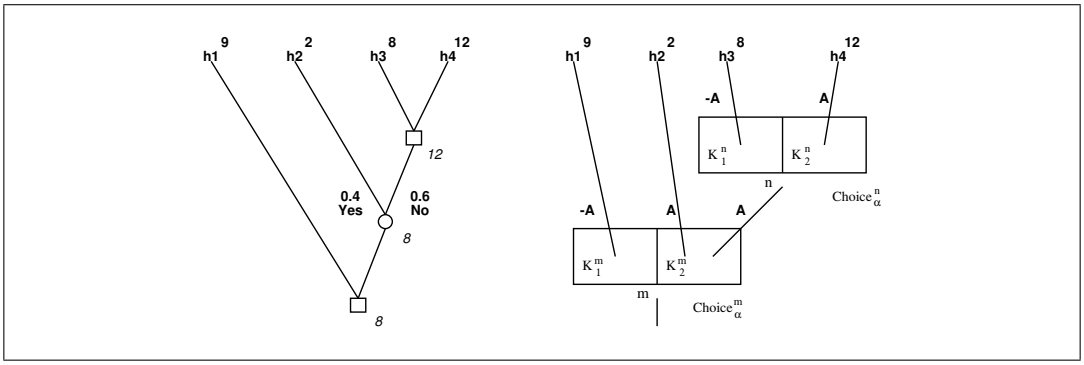


Figure 1: A decision tree and the corresponding utilitarian STIT-model

for 'the atomic proposition A', 'some time in the future A will be the case', 'agent α Sees To It That A', and 'it ought to be that A', respectively.

A is true at a moment-history pair m, h if and only if it is assigned the value true in the STIT-model, FA is true at a moment-history pair m, h if and only if there is some future moment on the history where A is true, $[\alpha\ cstit : A]$ is true at a moment history pair m, h if and only if A is true at all moment-history pairs through m that belong to the same *action* as m, h, and $\bigcirc A$ is true at a moment history pair m, h if and only if there is some history h' through m such that A is true at all pairs m, h'' for which the history h'' has a utility at least as high as h' ('moment determinate').

This semantic condition for the STIT-ought is a utilitarian generalisation of the standard deontic logic view (SDL) that 'it ought to be that A' means that A holds in all deontically optimal worlds.

On the STIT-model of Figure 1 we have $\mathcal{M}, m, h_3 \models A$ (directly from the valuation of atomic propositions on moment-history pairs), $\mathcal{M}, m, h_3 \models F\neg A$ (the proposition $\neg A$ is true later on, at moment n, on the history h_3 through m).

Also we have $\mathcal{M}, m, h_3 \models [\alpha\ cstit : A]$, because A holds for all histories through moment m belonging to the same action as h_3 (i.e. action K^m_2). Regarding ought-formulas we have: $\mathcal{M}, m, h_3 \models \bigcirc A$ and $\mathcal{M}, m, h_3 \models \bigcirc[\alpha\ cstit : A]$.

These two propositions are true for the same reason: the history h_4 through m has the highest utility (which means that we do not have to check conditions for histories with even higher utility) and satisfies both A and $[\alpha\ cstit : A]$ at m.

2.2 Gambling problem

Horty argues that ought-to-do statements are not just special kinds of ought-to-be statements. In particular, he claims that 'agent α ought to see to it that A' cannot

be modelled by the formula $\bigcirc[\alpha\ cstit : A]$ ('it ought to be that agent α sees to it that A').

Justification of this claim is found in the 'gambling example'. This example concerns the situation where an agent faces the choice between gambling to double or lose five dollar (action K_1) and refraining from gambling (action K_2). This situation is sketched in the figure below.

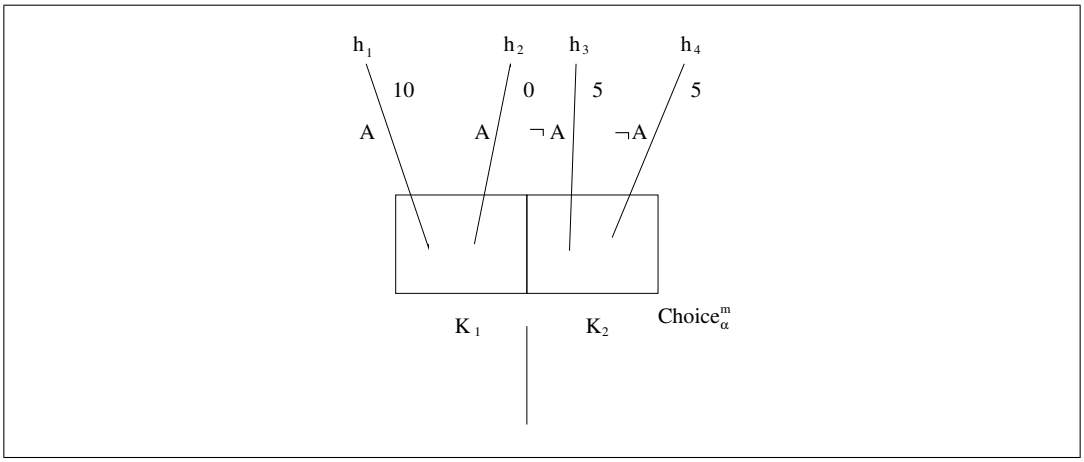


Figure 2: The gambling problem

The two histories that are possible by choosing action K_1 represent ending up with ten dollar by gaining five, and ending up with nothing by loosing all, respectively.

Also for action K_2, the game event causes histories to branch. But, for this action the two branches have equal utilities because the agent is not taking part in the game, thereby preserving his 5 dollar. Note this points to redundancy in the model representation: the two branches are logically indistinguishable, because there is no formula whose truth value would change by dropping one of them.

$\bigcirc[\alpha\ cstit : A]$ is true at m for history h_1 and for all histories with a higher utility (i.e. none), the formula $[\alpha\ cstit : A]$ is true. However, a reading of $\bigcirc[\alpha\ cstit : A]$ as 'agent α ought to perform action K_1' is counter-intuitive for this example. From the description of the gambling scenario it does not follow that one action is better than the other. In particular, without knowing the odds (the probabilities), we cannot say anything in favor of action K_1: by choosing it, we may either end up with more or with less utility than by doing K_2. The only thing one may observe is that action K_1 will be preferred by more adventurous agents. But that is not something the logic is concerned with.

This demonstrates that 'agent α ought to see to it that A' cannot be modelled by $\bigcirc[\alpha\ cstit : A]$. The cause of the mismatch can be explained as follows. Adapting and generalising the main idea behind SDL to the STIT-context, ought-to-be statements concern truth in a set of optimal histories ('worlds' in SDL). Optimality is directly determined by the utilities associated with individual histories. If ought-to-be is about optimal histories, then ought-to-do is about optimal actions. But, since actions are assumed to be non-deterministic, actions do not correspond with individual histories, but with *sets* of histories. This means that to apply the idea of optimality to the definition of ought-to-do operators, we have to generalise the notion of optimality such that it applies to *sets* of histories, namely, the sets that make up non-deterministic actions. More specifically, we have to *lift* the ordering of histories to an ordering of actions. The ordering of actions suggested by Horty is very simple: an action is strictly better than another action if all of its histories are at least as good as any history of the other action, and not the other way around.

Having lifted the ranking of histories to a ranking of actions, the utilitarian ought conditions can now be applied to actions. Thus, Horty defines the new operator 'agent α ought to see to it that A (in formula form: $\odot[\alpha\ cstit : A]$)' as the condition that for all actions not resulting in A there is a higher ranked action that does result in A, plus that all actions that are ranked even higher also result in A. This 'solves' the gambling problem. We do not have $\odot[\alpha\ cstit : A]$ or $\odot[\alpha\ cstit : \neg A]$ in the gambling scenario, because in the ordering of actions, K_1 is not better or worse than K_2.

3 Moral luck and the driving example

The gambling problem may be seen as a kind of moral luck: whether we obtain the utility of 10 or 0 is not due to our actions, but due to luck. The issue of moral luck is even more interesting in the case of multiple agents, where it depends on the actions of other agents whether you get utility 10 or 0.

Challenge 3. *How to deal with moral luck in normative reasoning?*

The driving example [58, p.119-121] is used to illustrate the difference between so-called dominance act utilitarianism and orthodox perspective on the agent's ought. Roughly, dominance act utilitarianism is that α ought to see to it that A just in case the truth of A is guaranteed by each of the optimal actions available to the agent—formally, that $\odot[\alpha\ cstit : A]$ should be settled true at a moment m just in case $K \subseteq |A|_m$ for each $K \in Optimal_\alpha^m$. When we adopt the orthodox perspective, the truth or falsity of ought statements can vary from index to index. The orthodox

perspective is that α should see to it that A at a certain index just in case the truth of A is guaranteed by each of the actions available to the agent that are optimal given the circumstances in which he finds himself at this index.

> "In this example, two drivers are travelling toward each other on a one-lane road, with no time to stop or communicate, and with a single moment at which each must choose, independently, either to swerve or to continue along the road. There is only one direction in which the drivers might swerve, and so a collision can be avoided only if one of the drivers swerves and the other does not; if neither swerves, or both do, a collision occurs. This example is depicted in Figure 3, where α and β represent the two drivers, K_1 and K_2 represent the actions available to α of swerving or staying on the road, K_3 and K_4 likewise represent the swerving or continuing actions available to β, and m represents the moment at which α and β must make their choice. The histories h_1 and h_3 are the ideal outcomes, resulting when one driver swerves and the other one does not; collision is avoided. The histories h_2 and h_4, resulting either when both drivers swerve or both continue along the road, represent non-ideal outcomes; collision occurs. The statement A, true at h_1 and h_2, expresses the proposition that α swerves." [58, p.119]

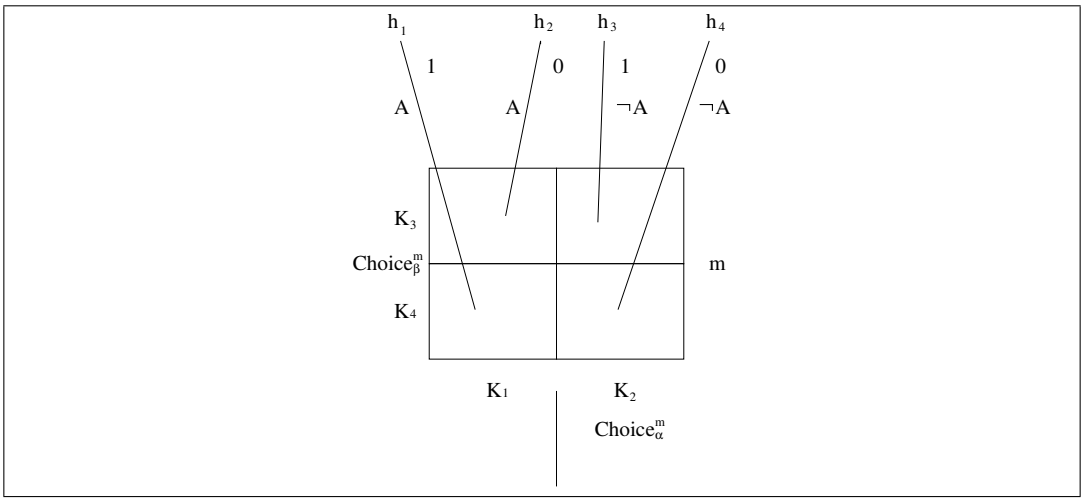


Figure 3: The driving example and moral luck

From the dominance point of view both actions available to α are classified as optimal, written as $Optimal_\alpha^m = \{K_1, K_2\}$. One of the optimal actions available to

α guarantees the truth of A and the other guarantees the truth of $\neg A$. Consequently $M, m \not\models \bigodot[\alpha \ cstit : A]$ and $M, m \not\models \bigodot[\alpha \ cstit : \neg A]$. From the orthodox point of view, we have $M, m, h_1 \models \bigcirc[\alpha \ cstit : A]$ and $M, m, h_2 \models \bigcirc[\alpha \ cstit : \neg A]$. What α ought to do at an index depends on what β does.

Horty concludes that from the standpoint of intuitive adequacy, the contrast between the orthodox and dominance deontic operators provides us with another perspective on the issue of moral luck, the role of external factors in our moral evaluations [58, p.121]. The orthodox ought is the one who after the actual event looks back to it. For example, when there has been a collision then α might say—perhaps while recovering from the hospital bed—that he ought to have swerved. The dominance ought is looking forward. Though the agent may legitimately regret his choice, it is not one for which he can be blamed, since either choice, at the time, could have led to a collision.

4 Procrastination: actualism vs possibilism

Practical reasoning is intimately related to reasoning about time. For example, if you are obliged and willing to visit a relative, but you always procrastinate this visit, then we may conclude that you violated this obligation. In other words, each obligation to do an action should come with a deadline [22, 11].

Challenge 4. *How to deal with procrastination in normative reasoning?*

The example of Procrastinate's choices [58, p. 162] illustrates the notion of strategic oughts. A strategy is a generalized action involving a series of actions. Like an action, a strategy determines a subset of histories. The set of admissible histories for a strategy σ is denoted $Adh(\sigma)$.

A crucial new concept here is the concept of a *Field*, which is basically a subtree of the STIT model which denotes that the agent's reasoning is limited to this range. A strategic ought is defined analogous to dominance act utilitarianism, in which action is replaced by strategy in a field. α ought to see to it that A just in case the truth of A is guaranteed by each of the optimal strategies available to the agent in the field—formally, that $\bigodot[\alpha \ cstit : A]$ should be settled true at a moment m just in case $Adh(\sigma) \subseteq |A|_m$ for each $\sigma \in Optimal^m_\alpha$. Horty observes some complications, and that a 'proper treatment of these issues might well push us beyond the borders of the current representational formalism' [p.150].

Horty also uses the example of Procrastinate's choices to distinguish between actualism and possibilism, for which he uses the strategic oughts, and in particular the notion of a field. Roughly, actualism is the view that an agent's current actions

are to be evaluated against the background of the actions he is actually going to perform in the future. Possibilism is the view that an agent's current actions are to be evaluated against the background of the actions that he might perform in the future, the available future actions.

The example is due to Jackson and Pargetter [63].

> "Professor Procrastinate receives an invitation to review a book. He is the best person to do the review, has the time, and so on. The best thing that can happen is that he says yes, and then writes the review when the book arrives. However, suppose it is further the case that were to say yes, he would not in fact get around to writing the review. Not because of incapacity or outside interference or anything like that, but because he would keep on putting the task off. (This has been known to happen.) This although the best thing that can happen is for Procrastinate to say yes and then write, and he *can* do exactly this, what *would* happen in fact were he to say yes is that he would not write the review. Moreover, we may suppose, this latter is the worst thing which may happen.
>
> [...]
>
> According to possibilism, the fact that Procrastinate would not write the review were he to say yes is irrelevant. What matters is simply what is possible for Procrastinate. He can say yes and then write; that is best; that requires *inter alia* that he says yes; therefore, he ought to say yes. According to actualism, the fact that Procrastinate would not actually write the review were he to say yes is crucial. It means that to say yes would be in fact to realize the worst. Therefore, Procrastinate ought to say no."

Horty represents the example by the STIT model in Figure 4. Here, m_1 is the moment at which Procrastinate, represented as the agent α, chooses whether or not to accept the invitation: K_1 represents the choice of accepting, K_2 the choice of declining. If Procrastinate accepts the invitation, he then faces at m_2 the later choice of writing the review or not: K_3 represents the choice of writing the review, K_4 another choice that results in the review not being written. For convenience, Horty also supposes that at m_3 Procrastinate has a similar choice whether or not to write the review: K_5 represents the choice of writing, K_6 the choice of not writing. The history h_1, in which Procrastinate accepts the invitation and then writes the review, carries the greatest value of 10; the history h_2, in which Procrastinate accepts the invitation and then neglects the task, the least value of 0; the history h_4, in which he declines, such that a less competent authority reviews the book, carries an inter-

mediate value of 5; and the peculiar h_3, in which he declines the invitation but then reviews the book anyway, carries a slightly lower value of 4, since he wastes his time, apart from doing no one else any good. The statement A represents the proposition that he accepts the invitation; the statement B represents the proposition that Procrastinate will write the review.

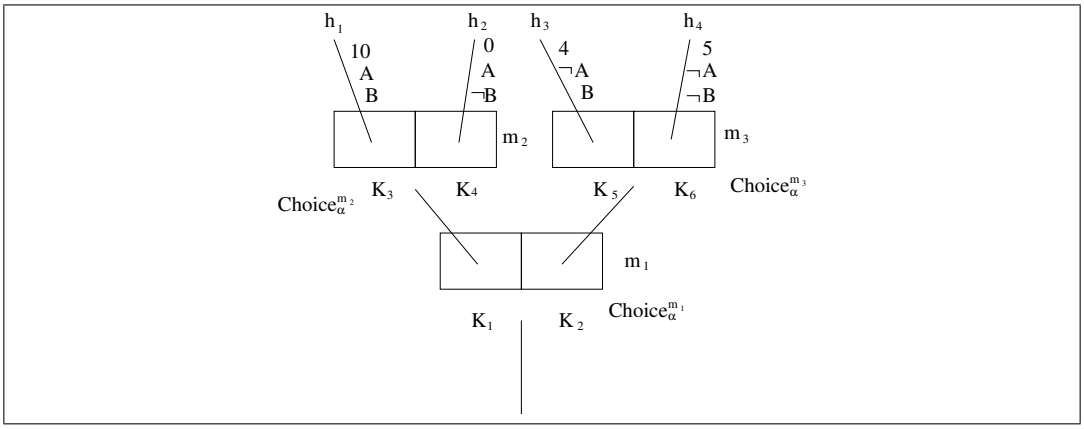


Figure 4: Procrastinate's choices

Now, in the possibilist interpretation, $M = \{m_1, m_2, m_3\}$ is the background field. In this interpretation, Procrastinate ought to accept the invitation because this is the action determined by the best available strategy—first accepting the invitation, and then writing the review. Formally, we have $Optimal_\alpha^M = \{\sigma_6\}$ with $\sigma_6 = \{\langle m_1, K_1\rangle, \langle m_2, K_3\rangle\}$. Since $Adh(\sigma_6) \subseteq |A|_m$, the strategic ought statement $\bigodot[\alpha\ cstit : A]$ is settled true in the field M. In the actualist interpretation, the background field may be narrowed to the set $M' = \{m_1\}$, which shifts from the strategic to the momentary theory of oughts. In this case, we have $\bigodot[\alpha\ cstit : A]$ is settled false. It is as if we choose to view Procrastinate as gambling on his own later choice in deciding whether to accept the invitation. However, from this perspective, this should not be viewed as a gamble; an important background assumption—and the reason that he should decline the invitation—is that he will not, in fact, write the review.

5 Jørgensen's dilemma and the problem of detachment

A philosophical problem that has had a major impact in the development of deontic logic is Jørgensen's dilemma. In a nutshell, given that norms cannot be true or false, the dilemma implies that deontic logic cannot be based on traditional truth func-

tional semantics. In particular, building on a tradition of Alchourrón and Bulygin in the seventies, Makinson [84] argues that norms need to be represented explicitly. SDL, DSDL and STIT logic represent logical relations between deontic operators, but they do not explicitly represent a distinction between norms and obligations. The explicit representation of norms is the basis of alternative semantics, that breaks with the idea of traditional semantics that norms and obligations have truth values, and most importantly, that discards the main technical and conceptual tool of traditional semantics, namely possible worlds. As an example, in this section we illustrate this alternative semantics using input/output logic.

5.1 Jørgensen's dilemma

While normative concepts are the subject of deontic logic, it is quite difficult to see how there can be a logic of such concepts at all. Norms like individual imperatives, promises, legal statutes, and moral standards are usually not viewed as being true or false. E.g. consider imperative or permissive expressions such as "John, leave the room!" and "Mary, you may enter now": they do not describe, but demand or allow a behavior on the part of John and Mary. Being non-descriptive, they cannot meaningfully be termed true or false. Lacking truth values, these expressions cannot—in the usual sense—be premise or conclusion in an inference, be termed consistent or contradictory, or be compounded by truth-functional operators. Hence, though there certainly exists a logical study of normative expressions and concepts, it seems there cannot be a logic of norms: this is Jørgensen's dilemma [65, 84].

Though norms are neither true nor false, one may state that *according to the norms*, something ought to be done or is permitted: the statements "John ought to leave the room" and"Mary is permitted to enter" are then true or false descriptions of the normative situation. Such statements are sometimes called normative statements, as distinguished from norms. To express principles such as the principle of conjunction: $O(p \wedge q) \leftrightarrow (Op \wedge Oq)$, with Boolean operators having truth-functional meaning at all places, deontic logic has resorted to interpreting its formulas Op, Fp, Pp not as representing norms, but as representing such normative statements. A possible logic of normative statements may then reflect logical properties of underlying norms—thus logic may have a "wider reach than truth", as Von Wright [124] famously stated.

Since the truth of normative statements depends on a normative situation, in the way in which the truth of the statement "John ought to leave the room" depends on whether some authority ordered John to leave the room or not, it seems that norms must be represented in a logical semantics that models such truth or falsity. However, semantics used to model the truth or falsity of normative statements mostly fail to

include norms. Standard deontic semantics evaluates deontic formulas with respect to sets of worlds, in which some are ideal or better than others—Ox is then defined to be true if x is true in all ideal or the best reachable worlds. Alternatively, norms, not ideality, should provide the basis on which normative statements are evaluated. Thus the following question arises, asked by D. Makinson [84]:

Challenge 5. *How can deontic logic be reconstructed in accord with the philosophical position that norms are neither true nor false?*

In the older literature on deontic logic there has been a veritable 'imperativist tradition' of authors that have, deviating from the standard approach, in one way or other, tried to give truth definitions for deontic operators with respect to given sets of norms. Cf. among others S. Kanger [67], E. Stenius [105], T. J. Smiley [103], Z. Ziemba [125], B. van Fraassen [114], Alchourrón and Bulygin [2] and I. Niiniluoto [90]. The reconstruction of deontic logic as logic about imperatives has been the project of Jörg Hansen beginning with [47]. Input/output logic [85] is another reconstruction of a logic of norms in accord with the philosophical position that norms direct rather than describe, and are neither true nor false. We explain it in more detail in the next section below.

5.2 Input/output logic

To illustrate a possible answer to the dilemma, we use Makinson and van der Torre's input/output logic [85, 86, 87], and we therefore assume familiarity with this approach (cf. [88] for an introduction). Input/output logic takes a very general view at the process used to obtain conclusions (more generally: outputs) from given sets of premises (more generally: inputs). While the transformation may work in the usual way, as an 'inference motor' to provide logical conclusions from a given set of premises, it might also be put to other, perhaps non-logical uses. Logic then acts as a kind of secretarial assistant, helping to prepare the inputs before they go into the machine, unpacking outputs as they emerge, and, less obviously, coordinating the two. The process as a whole is one of logically assisted transformation, and is an inference only when the central transformation is so. This is the general perspective underlying input/output logic. It is one of logic at work rather than logic in isolation; not some kind of non-classical logic, but a way of using the classical one.

Suppose that we have a set G (meant to be a set of conditional norms), and a set A of formulas (meant to be a set of given facts). The problem is then: how may we reasonably define the set of propositions x making up the output of G given A, which we write $out(G, A)$? In particular, if we view the output as a collection of descriptions of states of affairs that ought to obtain given the norms G and the

facts A, what is a reasonable output operation that enables us to define a deontic O-operator that returns the normative statements that are true given the norms and the facts—the normative consequences given the situation? One such definition is the following:

$$G, A \models Ox \quad \text{iff} \quad x \in out(G, A)$$

So Ox is true iff the output of G under A includes x. Note that this is rather a description of how we think such an output should or might be interpreted, whereas 'pure' input/output logic does not discuss such definitions. For a simple case, let G include a conditional norm that states that if a is the case, x should obtain (we write $(a, x) \in G$). An unconditional norm that commits the agent to realizing x is represented by a conditional norm $(\top, x)$, where $\top$ means an arbitrary tautology. If a can be inferred from A, i.e. if $a \in Cn(A)$, and z is logically implied by x, then z should be among the normative consequences of G given A. An operation that does this is simple-minded output out_1:

$$out_1(G, A) \quad = \quad Cn(G(Cn(A)))$$

where $G(B) = \{y \mid (b, y) \in G \text{ and } b \in B\}$. So in the given example, Oz is true given $(a, x) \in G$, $a \in Cn(A)$ and $z \in Cn(x)$.

Simple-minded output may, however, not be strong enough. Sometimes, legal argumentation supports reasoning by cases: if there is a conditional norm (a, x) that states that an agent must bring about x if a is the case, and a norm (b, x) that states that the same agent must also bring about x if b is the case, and $a \vee b$ is implied by the facts, then we should be able to conclude that the agent must bring about x. An operation that supports such reasoning is basic output out_2:

$$out_2(G, A) \quad = \quad \cap\{Cn(G(V)) \mid v(A) = 1\}$$

where v ranges over Boolean valuations plus the function that puts $v(b) = 1$ for all formulae b, and $V = \{b \mid v(b) = 1\}$. It can easily be seen that now Ox is true given $\{(a, x), (b, x)\} \subseteq G$ and $a \vee b \in Cn(A)$.

This definition of out_2 may give rise to a mere feeling of merely technical adequacy, because of its recourse to intersection and valuations, neither of which quite corresponds to our natural course of reasoning in such situations. However, this semantics makes explicit what is present but implicit in the use of possible worlds in conditional logics: if you want to reason by cases in the logic, you need to represent the cases explicitly in the semantics.

It is quite controversial whether reasoning with conditional norms should support 'normative' or 'deontic detachment', i.e. whether it should be accepted that if one norm (a, x) commands an agent to make x true in conditions a, and another norm (x, y) directs the agent to make y true given x is true, then the agent has an obligation to make y true if a is factually true. Some would argue that as long as the agent

has not in fact realized x, the norm to bring about y is not 'triggered'; others would maintain that obviously the agent has an obligation to make $x \wedge y$ true given that a is true. Moreover, the inference can be restricted to cases where the agent ought to make x true instantly rather than eventually, see [84, 11] If such detachment is viewed as permissible for normative reasoning, then one might use reusable output out_3 that supports such reasoning:

$$out_3(G, A) \quad = \quad \cap\{Cn(G(B)) \mid A \subseteq B = Cn(B) \supseteq G(B)\}$$

An operation that combines reasoning by cases with deontic detachment is then reusable basic output out_4:

$$out_4(G, A) \quad = \quad \cap\{Cn(G(V)) : v(A) = 1 \text{ and } G(V) \subseteq V\}$$

It may turn out that further modifications of the output operation are required in order to produce reasonable results for normative reasoning. Also, the proposal to employ input/output logic to reconstruct deontic logic may lead to competing solutions, depending on what philosophical views as to what transformations should be acceptable one subscribes to. All this is what input/output logic is about. However, it should be noted that input/output logic succeeds in representing norms as entities that are neither true nor false, while still permitting normative reasoning about such entities.

5.3 Contrary to duty reasoning reconsidered

In the input/output logic framework, the strategy for eliminating excess output is to cut back the set of generators to just below the threshold of yielding excess. To do that, input/output logic looks at the maximal non-excessive subsets, as described by the following definition:

Definition (Maxfamilies) *Let G be a set of conditional norms and A and C two sets of propositional formulas. Then maxfamily(G, A, C) is the set of maximal subsets $H \subseteq G$ such that out$(H, A) \cup C$ is consistent.*

For a possible solution to Chisholm's paradox, consider the following output operation $out^{\cap}$:

$$out^{\cap}(G, A) \quad = \quad \bigcap\{out(H, A) \mid H \in maxfamily(G, A, A)\}$$

So an output x is in $out^{\cap}(G, A)$ if it is in output $out(H, A)$ of all maximal norm subsets $H \subseteq G$ such that $out(H, A)$ is consistent with the input A. Let a deontic O-operator be defined in the usual way with regard to this output:

$$G, A \models O^{\cap}x \quad \text{iff} \quad x \in out^{\cap}(G, A)$$

Furthermore, tentatively, and only for the task of shedding light on Chisholm's paradox, let us define an entailment relation between norms as follows:

Definition (Entailment relation) *Let G be a set of conditional norms, and (a, x) be a norm whose addition to G is under consideration. Then (a, x) is entailed by G iff for all sets of propositions A, $out^{\cap}(G \cup \{(a, x)\}, A) = out^{\cap}(G, A)$.*

So a (considered) norm is entailed by a (given) set of norms if its addition to this set would not make a difference for any set of facts A. Finally, let us use the following cautious definition of 'coherence from the start' (also called 'minimal coherence' or 'coherence per se'), see Section 7:

A set of norms G is 'coherent from the start' iff $\bot \notin out(G, \top)$.

Now consider a 'Chisholm norm set' $G = \{(\top, x), (x, z), (\neg x, \neg z), \}$, where $(\top, x)$ means the norm that the man must go to the assistance of his neighbors, (x, z) means the norm that it ought to be that if he goes he ought to tell them he is coming, and $(\neg x, \neg z)$ means the norm that if he does not go he ought not to tell them he is coming. It can be easily verified that the norm set G is 'coherent from the start' for all standard output operations out_n, since for these either $out(G, \top) = Cn(\{x\})$ or $out(G, \top) = Cn(\{x, z\})$, and both sets $\{x\}$ and $\{x, z\}$ are consistent. Furthermore, it should be noted that all norms in the norm set G are independent from each other, in the sense that no norm $(a, x) \in G$ is entailed by $G \setminus \{(a, x)\}$ for any standard output operation $out_n^{(+)}$: for $(\top, x)$ we have $x \in out^{\cap}(G, \top)$ but $x \notin out^{\cap}(G \setminus \{(\top, x)\}, \top)$, for (x, z) we have $z \in out^{\cap}(G, x)$ but $z \notin out^{\cap}(G \setminus \{(x, z)\}, x)$, and for $(\neg x, \neg z)$ we have $\neg z \in out^{\cap}(G, \neg x)$ but $\neg z \notin out^{\cap}(G \setminus \{(\neg x, \neg z)\}, \top)$. Finally consider the 'Chisholm fact set' $A = \{\neg x\}$, that includes as an assumed unalterable fact the proposition $\neg x$, that the man will not go to the assistance of his neighbors: we have $maxfamily(G, A, A) = \{G \setminus \{(\top, x)\}\} = \{\{(x, z), (\neg x, \neg z), \}\}$ and either $out(G \setminus \{(\top, x)\}, A) = Cn(\{\neg z\})$ or $out(G \setminus \{(\top, x)\}, A) = Cn(\{\neg x, \neg z\})$ for all standard output operations $out_n^{(+)}$, and so $O^{\cap}\neg z$ is true given the norm and fact sets G and A, i.e. the man must not tell his neighbors he is coming. Thus:

$G, A \models O^{\cap}\neg z$

6 Multiagent detachment

In Section 6.1 we introduce normative multiagent systems using agents and controllable propositions, and we introduce a challenge for detachment for multiagent systems. In Section 6.2 we give a solution for the challenge in these formalisms.

6.1 Challenge for multiagent detachment

Olde Loohuis [91] argues that the assumption that other agents comply with their norms reflects that agents live in a responsible world. However, Makinson [84]

observes that if all we know is that "John owes Peter $1000" and "if John pays Peter $1000, then Peter is obliged to give John a receipt," then we cannot detach that Peter has to give John a receipt unconditionally based on the assumption that John will pay Peter the money.

We assume that the normative system is known to all agents, and in this section we assume that it does not change over time, and that each norm is directed to one agent only. The agents reason about the consequences of the normative system, that is, which obligations and permissions can be detached from it. With an explicit normative system, the agents should act such that they do not violate norms. Moreover, in this section we assume that each (instance of a) norm specifies the behavior of a single individual agent. For example, a norm may say that an agent should drive to the right hand side of the street, but we do not consider group norms saying that agents should live together in harmony.

We do not assume a full action theory as in STIT logic, but we assume a minimal action theory: the set of propositions is partitioned into parameters (uncontrollable propositions) and decision variables (controllable propositions). Boutilier [19] traces this idea back to discrete event systems, see also Cholvy and Garion [30]. It is an abstract and general approach, since we can instantiate the propositions with action descriptions like do(action) or done(action). Note that this generality is in line with game theory, which abstracts away sequential decisions in extensive games by representing conditional plans as strategic games. Boutilier observes that the theory can be extended to a full fledged action theory by, for example, introducing a causal theory. By convention, the proposition letters p, p_1, etc are parameters, a, a_1, ..., are decision variables for agent 1, b, b_1, ..., are decision variables for agent 2, etc. Norms are written as pairs of propositional formulas, where (p_1, p_2) is read as "if p_1 is the case, then p_2 ought to be the case," (a_1, a_2) is read as "if agent 1 does a_1, then he has to do a_2," and so on. We restrict the propositional language to conjunctions of literals (propositional atoms or their negations), so we do not consider disjunctions or material implications.

Definition 6.1 (Normative multi agent system, individual norms)**.** *A normative multiagent system is a tuple NMAS= $\langle A, P, c, N\rangle$ where A is a set of agents, P is a set of atomic propositions, $c : P \to A$ is a partial function which maps the propositions to the agents controlling them, and N is a set of pairs of conjunctions of literals built of P, such that if $(\phi, \psi) \in N$, then all propositional atoms in ψ are controlled by a single agent.*

Our action theory may be seen as a simple kind of STIT theory, in the sense that an obligation for a proposition p controlled by agent α may be read as: "the agent α ought to see to it that p is the case." Though this abstracts away from the

temporal issues of STIT operators, it still has the characteristic property of STIT logics that actions have a higher granularity than worlds.

Makinson [84] illustrates the intricacies of temporal reasoning with norms, obligations and agents by discussing the iteration of detachment, in the sense that from the two conditional norms "if ϕ, then obligatory ψ" and "if ψ, then obligatory χ" together with the fact ϕ, we can derive not only that ψ is obligatory, but also that χ is obligatory. Makinson's challenge is how to detach obligations based on the principle that agents cannot assume that other agents comply with their norms, but they assume that they themselves comply with their norms. In other words, deontic detachment holds only for the single agent a-temporal case.

First, Makinson argues that iteration of detachment often appears to be appropriate. He gives the following example, based on instructions to authors preparing manuscripts.

Example 6.2 (Manuscript [84])**.** *Let the set of norms be as follows:* $(25x15, 12)$*="if* $25x15$*, then obligatory* 12*"* *and* $(12, refs10)$*="if* 12*, then obligatory* $refs10$*", where* $25x15$ *is "The text area is 25 by 15 cm",* 12 *is "The font size for the main text is 12 points", and* $refs10$ *is "The font size for the list of references is 10 points". Moreover, consider a single agent controlling the three variables. If the facts contain* $25x15$*, then we want to detach not only that it is obligatory that* 12*, but also that it is obligatory that* $refs10$*.*

Second, he argues that iteration of detachment sometimes appears to be inappropriate by discussing the following example, which he attributes to Sven Ove Hansson.

Example 6.3 (Receipt [84])**.** *Let instances of the norms be*
(owe_{jp}, pay_{jp})*="if* owe_{jp}*, then obligatory* pay_{jp}*"* *and*
$(pay_{jp}, receipt_{pj})$*="if* pay_{xy}*, then obligatory* $receipt_{pj}$*"*
where owe_{jp} *is "John owes Peter \$1000",* pay_{jp} *is "John pays Peter \$1000", and* $receipt_{pj}$ *is "Peter gives John a receipt for \$1000". Moreover, assume that the first variable is not controlled by an agent, the second is controlled by John, and the third is controlled by Peter. Intuitively Makinson would say that in the circumstance that John owes Peter \$1000, considered alone, Peter has no obligation to write any receipt. That obligation arises only when John fulfils his obligation.*

Makinson observes that there appear to be two principal sources of difficulty here. One concerns the passage of time, and the other concerns bearers of the obligations. Sven Ove Hansson's example above involves both of these factors.

> "We recall that our representation of norms abstracts entirely from the question of time. Evidently, this is a major limitation of scope, and leads

> to discrepancies with real-life examples, where there is almost always an implicit time element. This may be transitive, as when we say "when b holds then a should eventually hold", or "... should simultaneously hold". But it may be intransitive, as when we say "when b holds then a should hold within a short time" or "... should be treated as a matter of first priority to bring about". Clearly, iteration of detachment can be legitimate only when the implicit time element is either nil or transitive. Our representation also abstracts from the question of bearer, that is, who (if anyone) is assigned responsibility for carrying out what is required. This too can lead to discrepancies. Iteration of detachment becomes questionable as soon as some promulgations have different bearers from others, or some are impersonal (i.e. without bearer) while others are not. Only when the locus of responsibility is held constant can such an operation take place." [84]

Challenge 6. *How to define detachment for multiple agents?*

Broersen and van der Torre [21] consider the temporal aspects of the example. In this section we consider the actions of the agents. The following example extends the discussion of the example to aggregative deontic detachment.

Example 6.4 (continued)**.** *Consider again* (owe_{jp}, pay_{jp}) *and* $(pay_{jp}, receipt_{pj})$*, where the first variable is not controlled by an agent, the second is controlled by John, and the third is controlled by Peter. In the circumstance that John owes Peter \$1000, considered alone, do we want to derive the obligation for* $pay_{jp} \wedge receipt_{pj}$*, that is, the obligation that "John pays Peter \$1000", and "Peter gives John a receipt for \$1000"? In many systems the obligation for* $pay_{jp} \wedge receipt_{pj}$ *implies the obligation for* $receipt_{pj}$*, such that the answer will be negative. However, if the obligation for* $pay_{jp} \wedge receipt_{pj}$ *does not imply the obligation for* $receipt_{pj}$*, then maybe the obligation for* $pay_{jp} \wedge receipt_{pj}$ *is not as problematic as the obligation for* $receipt_{pj}$*. Moreover, the obligation for* $pay_{jp} \wedge receipt_{pj}$ *is a compact representation of the fact that ideally, the exchange of money and receipt takes place.*

6.2 Deontic detachment for agents

As the iterative approaches seem most natural to most people, we define deontic detachment of agents using these iterative approaches. The question thus arises whether we consider sequential or iterated detachment. The following example illustrates this question, not discussed by Makinson [84].

Example 6.5. *$N = \{(p, a), (a, b_1), (a \wedge b_1, b_2)\}$ where p is a parameter, a is a decision variable of agent 1, and b_1 and b_2 are decision variables of agent 2. In context $F = \{p, a\}$, do we want to detach only b_1, or both b_1 and b_2? If we can detach b_2, then this implies that despite the fact that a and b_1 are decision variable from distinct agents we can use $(a \wedge b_1, b_2)$ to detach b_2.*

In the above example, we believe that b_2 should be derivable, because only b_1 is reused when b_2 is detached, and both b_1 and b_2 are decision variables of the same agent. In other words, when considering the norm $(a \wedge b_1, b_2)$ to detach b_2, we should not consider the norm and reject it because there is a variable in the input which refers to another agent, but we should consider it since we have $a \in F$ as a fact, and b_1 already in the output, we can derive b_2 too.

If b_2 should not be derivable, then we could simply restrict the set of norms that we select from N to satisfy the syntactic criterion, just like we selected the set of norms N_0. However, if b_2 should be derivable, then we have to define detachment procedures for each agent, and combine them afterwards. This is formalized in the following detachment procedure for agents.

Definition 6.6 (Iterative detachment for agents.)**.** *Agent $a \in A$ controls a propositional formula ϕ, written as $c(\phi) = a$, if and only if for all atoms $x \in \phi$ we have $c(x) = a$.*

$$N_0^a = \{(\phi, \psi) \in N \mid F \cup \{\phi\} \not\models \neg\psi, c(\psi) = a\}$$

$E_0^{ia} = \emptyset$. For $n = 1$ to ∞ do $E_{n+1}^{ia} = \{\psi \mid (\phi, \psi) \in N_0^a, F \cup E_n^{ia} \models \phi\}$ if consistent with F, E_n^{ia} otherwise. $out^{ia}(N, F, a) = Cn(\cup E_i^{ia})$, and $out^{ia}(N, F) = \cup_{a \in A} out^{ia}(N, F, a)$.

We leave the logical analysis of this ans related approaches to future work.

7 Coherence

Consider norms which at the same time require you to leave the room and not to leave the room. In such cases, we are inclined to say that there is something wrong with the normative system. This intuition is captured by the SDL axiom $D : \neg(Ox \wedge O\neg x)$ that states that there cannot be co-existing obligations to bring about x and to bring about $\neg x$, or, using the standard cross-definitions of the deontic modalities: x cannot be both, obligatory and forbidden, or: if x is obligatory then it is also permitted. However, what does this tell us about the normative system?

Since norms do not bear truth values, we cannot, in any usual sense, say that such a set of norms is inconsistent. All we can consider is the consistency of the output of a set of norms. We like to use the term *coherence* with respect to a set of

norms with consistent output. For a start, consider the notion of minimal coherence in Section 5.3:

(0) A set of norms G is minimal coherent iff $\bot \notin out(G, \emptyset)$.

This is clearly very weak, as for example the norms $(a, x), (a, \neg x)$ would be coherent. Alternatively, we might try to define coherence as follows:

(1) A set of norms G is coherent iff $\bot \notin out(G, A)$.

However, this definition seems not quite sufficient: one might argue that one should be able to determine whether a set of norms G is coherent or not regardless of what arbitrary facts A might be assumed. A better definition would be (1a):

(1a) A set of norms G is coherent iff there exists a set of formulas A such that $\bot \notin out(G, A)$.

For (1a) it suffices that there exists a situation in which the norms can be, or could have been, fulfilled. However, consider the set of norms $G = \{(a, x), (a, \neg x)\}$ that requires both x to be realized and $\neg x$ to be realized in conditions a: it is immediate that e.g. for all output operations out_n, we have $\bot \notin out_n(G, \neg a)$: no conflicting demands arise when $\neg a$ is factually assumed. Yet something seems wrong with a normative system that explicitly considers a fact a only to tie to it conflicting normative consequences. The dual of (1a) would be

(1b) A set of norms G is coherent iff for all sets of formulas A, we have $\bot \notin out(G, A)$.

Now a set G with $G = \{(a, x), (a, \neg x)\}$ would no longer be termed coherent. (1b) makes the claim that for no situation A, two norms $(a, x), (b, y)$ would ever come into conflict, which might seem too strong. We may wish to restrict A to sets of facts that are consistent, or that are not in violation of the norms. The question is, basically, how to distinguish situations that the norm-givers should have taken care of, from those that describe misfortune or otherwise unhappy circumstances. A weaker claim than (1b) would be (1c):

(1c) A set of norms G is coherent iff for all a with $(a, x) \in G$, we have $\bot \notin out(G, a)$.

By this change, consistency of output is required just for those factual situations that the norm-givers have foreseen, in the sense that they have explicitly tied normative consequences to such facts. Still, (1c) might require further modification, since if a is a foreseen situation, and so is b, then also $a \vee b$ or $a \wedge b$ might be counted as foreseen situations for which the norms should be coherent.

As one anonymous reviewer suggested, another solution consists in combining elements of previous proposals:

(1d) A set of norms G is coherent iff for each $A \subseteq \{a \mid (a, x) \in G\}$, if A is non-empty and consistent, then $\bot \notin out(G, A)$.

However, there is a further difficulty: let G contain a norm $(a, \neg a)$ that, for conditions in which a is unalterably true, demands that $\neg a$ be realized. We then have $\neg a \in out_n(G, a)$ for the principal output operations out_n, but not $\bot \in out_n(G, a)$. Certainly the term 'incoherent' should apply to a normative system that requires the agent to accomplish what is—given the facts in which the duty arises—impossible. However, since not every output operation supports 'throughput', i.e. the input is not necessarily included in the output, neither (1) nor its variants implies that the agent can actually realize all propositions in the output, though they might be logically consistent. We might therefore demand that the output be not merely consistent, but consistent with the input:

(2) A set of norms G is coherent iff $out(G, A) \cup A \not\models \bot$.

However, with definition (2) we obtain the questionable result that for any case of norm-violation, i.e. for any case in which $(a, x) \in G$ and $(a \wedge \neg x) \in Cn(A)$, G must be termed incoherent—Adam's fall would only indicate that there was something wrong with God's commands. One remedy would be to leave aside all those norms whose violation is entailed by the circumstances A, i.e. instead of $out(G, A)$ consider $out(\{(a, x) \in G \mid (a \wedge \neg x) \notin Cn(A)\}, A)$—but then a set G such that $(a, \neg a) \in G$ would not be incoherent.[2] It seems it is time to formally state our problem:

Challenge 7. *When is a set of norms to be termed 'coherent'?*

As can be seen from the discussion above, input/output logic provides the tools to formally discuss this question, by rephrasing the question of coherence of the norms as one of consistency of output, and of output with input. Both notions have been explored in the input/output framework as 'output under constraints', see also the motivation regarding contrary-to-duty reasoning in Section 1.4.:

Definition (Output under constraints) *Let G be a set of conditional norms and A and C two sets of propositional formulas. Then G is coherent in A under constraints C when $out(G, A) \cup C$ is consistent.*

Future study must define an output operation, determine the relevant states A, and find the constraints C, such that any set of norms G would be appropriately termed coherent or incoherent by this definition.

[2]Temporal dimensions are not considered here. In an approach that would consider dynamic norms, one may argue, throughput should not be included in a definition of coherence as any change involves an inconsistency between the way things were and the way they become.

8 Normative conflicts and dilemmas

There are essentially two views on the question of normative conflicts: in the one view, they do not exist. In the other view, conflicts and dilemmas are ubiquitous.

According to the view that normative conflicts are ubiquitous, it is obvious that we may become the addressees of conflicting normative demands at any time. My mother may want me to stay inside while my brother wants me to go outside with him and play games. I may have promised to finish a paper by the end of a certain day, while for the same day I have promised a friend to come to dinner—now it is late afternoon and I realize I will not be able to finish the paper if I visit my friend. Social convention may require me to offer you a cigarette when I am lighting one for myself, while concerns for your health should make me not offer you one. Legal obligations might collide - think of the case where the SWIFT international money transfer program was required by US anti-terror laws to disclose certain information about its customers, while under European law that also applied to that company, it was required not to disclose this information. Formally, let there be two conditional norms (a, x) and (b, y): unless we have that either $(x \to y) \in Cn(a \wedge b)$ or $(y \to x) \in Cn(a \wedge b)$ there is a possible situation $a \wedge b \wedge \neg(x \wedge y)$ in which the agent can still satisfy each norm individually, but not both norms collectively. But to assume this for any two norms (a, x) and (b, y) is clearly absurd. Nevertheless, as discussed extensively in Section 1 of this article, Lewis's [74, 75] and Hansson's [53] deontic semantics imply that there exists a 'system of spheres', in our setting: a sequence of boxed contrary-to-duty norms $(\top, x_1), (\neg x_1, x_2), (\neg x_1 \wedge \neg x_2, x_3), ...$ that satisfies this condition. So any logic about norms must take into account possible conflicts. But standard deontic logic SDL includes D: $\neg(Ox \wedge O\neg x)$ as one of its axioms, and it is not immediately clear how deontic reasoning could accommodate conflicting norms.

Challenge 8a. *How can deontic logic accommodate possible conflicts of norms?*

The literature on normative conflicts and dilemmas is vast. As highlighted earlier in this article, here we do not aim at an exhausting literature review on the topic; for that, the interested reader is referred to Goble's [38] chapter in the handbook of deontic logic and normative systems. If we accept the view that normative conflicts not only genuinely exist but are also ubiquitous, one classical way to deal with such conflicts consists in denying that 'ought' implies 'can', as done by Lemmon [73]. Another common solution is to deny the principle of conjunction, that is, to deny that oughting to do x and y separately implies ought to do both [89, 114, 35]. However, this solution was challenged by Horty's example [59, 60, 61, 62] where, from "Smith ought to fight in the army or perform alternative national service" and "Smith ought not to fight in the army", we should be able to derive "Smith

ought to perform alternative national service". By withdrawing the principle of conjunction, this argument is no longer valid. The distribution rule states that x necessitates y implies that, if one ought to do x, then one ought to do y. As Goble [38] observes, although this principle has been often criticized for its role in many deontic paradoxes, its responsibility in connection with normative conflicts has rarely been discussed. Keeping the principle of conjunction while removing the distribution rule would validate Horty's argument [37]. For other systems that restrict the distribution principle, see [36, 37].

In an input/output setting one could say that there exists a conflict whenever $\bot \in Cn(out(G, A) \cup A)$, i.e. whenever the output is inconsistent with the input: then the norms cannot all be satisfied in the given situation. There appear to be two ways to proceed when such inconsistencies cannot be ruled out. For the concepts underlying the 'some-things-considered' and 'all-things-considered' O-operators defined below cf. Horty [60] and Hansen [48, 49]. For both, it is necessary to recur to the the notion of a $maxfamily(G, A, A)$, i.e. the family of all maximal $H \subseteq G$ such that $out(H, A) \cup A$ is consistent. On this basis, input/output logic defines the following two output operations $out^{\cup}$ and $out^{\cap}$:

$$\begin{aligned} out^{\cup}(G, A) &= \bigcup\{out(H, A) \mid H \in maxfamily(G, A, A)\} \\ out^{\cap}(G, A) &= \bigcap\{out(H, A) \mid H \in maxfamily(G, A, A)\} \end{aligned}$$

Note that $out^{\cup}$ is a non-standard output operation that is not closed under consequences, i.e. we do not generally have $Cn(out^{\cup}(G, A)) = out^{\cup}(G, A)$. Finally we may use the intended definition of an O-operator

$$G, A \models Ox \quad \text{iff} \quad x \in out(G, A)$$

to refer to the operations $out^{\cup}$ and $out^{\cap}$, rather than the underlying operation $out(G, A)$ itself, and write $O^{\cup}x$ and $O^{\cap}x$ to mean that $x \in out^{\cup}(G, A)$ and $x \in out^{\cap}(G, A)$, respectively. Then we have that the 'some-things-considered', or 'bold' O-operator $O^{\cup}$ describes x as obligatory given the set of norms G and the facts A if x is in the output of some $H \in maxfamily(G, A, A)$, i.e. if some subset of non-conflicting norms, or: some coherent normative standard embedded in the norms, requires x to be true. It is immediate that neither the SDL axiom $D : \neg(Ox \wedge O\neg x)$ nor the agglomeration principle $C : Ox \wedge Oy \rightarrow O(x \wedge y)$ holds for $O^{\cup}$, as there may be two competing standards demanding x and $\neg x$ to be realized, while there may be none that demands the impossible $x \wedge \neg x$. However, the 'all-things-considered', or 'sceptic', O-operator $O^{\cap}$ describes x as obligatory given the norms G and the facts A if x is in the outputs of all $H \in maxfamily(G, A, A)$, i.e. it requires that x must be realized according to all coherent normative standards. Note that by this definition, both SDL theorems D and C are validated.

The opposite view, that normative conflicts do not exist, appeals to the very

notion of obligation: it is essential for the function of norms—to direct human behavior—that the subject of the norms is capable of following them. To state a norm that cannot be fulfilled is a meaningless use of language. To state two norms which cannot both be fulfilled is confusing the subject, not giving him or her directions. To say that a subject has two conflicting obligations is therefore a misuse of the term 'obligation'. So there cannot be conflicting obligations, and if things appear differently, a careful inspection of the normative situation is required that resolves the dilemma in favor of the one or other of what only appeared both to be obligations. In particular, this inspection may reveal that the apparent conflicts in reality comes from some ambiguities in the examples, for instance where a moral 'ought' is not compatible with a legal 'ought': thus, there is no real conflict, because the two 'oughts' refer to two different spheres, and each should be represented with a different operator [26, 27]. Or again, a priority ordering of the apparent obligations may help resolving the conflict, e.g. in Ross [100], von Wright [121, 122], and Hare [55]. The problem that arises for such a view is then how to determine the 'actual obligations' in face of apparent conflicts, or, put differently, in the face of conflicting 'prima facie' obligations.

Challenge 8b. How can the resolution of apparent conflicts be semantically modeled?

Again, both the $O^{\cup}$ and the $O^{\cap}$-operator may help to formulate and solve the problem: $O^{\cup}$ names the conflicting *prima facie* obligations that arise from a set of norms G in a given situation A, whereas $O^{\cap}$ resolves the conflict by only telling the agent to do what is required by all maximal coherent subsets of the norms: so there might be conflicting 'prima facie' $O^{\cup}$-obligations, but no conflicting 'all things considered' $O^{\cap}$-obligations. The view that a priority ordering helps to resolve conflicts seems more difficult to model. A good approach appears to be to let the priorities help us to select a set $P(G, A, A)$ of preferred maximal subsets $H \in \mathit{maxfamily}(G, A, A)$. We may then define the $O^{\cap}$-operator not with respect to the whole of $\mathit{maxfamily}(G, A, A)$, but only with respect to its selected preferred subsets $P(G, A, A)$. Ideally, in order to resolve all conflicts, the priority ordering should narrow down the selected sets to $\mathit{card}(P(G, A, A)) = 1$, but this generally requires a strict ordering of the norms in G. The demand that all norms can be strictly ordered is itself subject of philosophical dispute. Some moral requirements may be incomparable: this is Sartre's paradox, where the requirement that Sartre's student stays with his ailing mother conflicts with the requirement that the student joins the resistance against the German occupation [101]. Other moral requirements may be of equal weight, e.g. two simultaneously obtained obligations towards identical twins, of which only one can be fulfilled [89]. The difficult part is then to define a mechanism that determines the preferred maximal subsets by use of the given

priorities between the norms. There have been several proposals to this effect, not all of them successful, and the reader is referred to the discussions in Boella and van der Torre [13] and Hansen [50, 51].

9 Descriptive dyadic obligations

Dyadic deontic operators, that formalize e.g. 'x ought to be true under conditions a' as $O(x|a)$, were introduced over 50 years ago by G. H. von Wright [118]. Their introduction was due to Prior's paradox of derived obligation: often a primary obligation Ox is accompanied by a secondary, 'contrary-to-duty' obligation that pronounces y (a sanction, a remedy) as obligatory if the primary obligation is violated. At the time, the usual formalization of the secondary obligation would have been $O(\neg x \to y)$, but given Ox and the axioms of standard deontic logic SDL, $O(\neg x \to y)$ is derivable for any y. A bit later, Chisholm's paradox showed that formalizing the secondary obligation as $\neg x \to Oy$ produces similarly counterintuitive results. So to deal with such contrary-to-duty conditions, the dyadic deontic operator $O(x|a)$ was invented. For a historical account the reader is referred to Hilpinen and McNamara's chapter in the handbook of deontic logic and normative systems [57].

In Section 1.3 we have extensively discussed DSDL. The perhaps best-known semantic characterization of dyadic deontic logic is B. Hansson's [53] system DSDL3, axiomatized by Spohn [104]. Hansson's idea was that the circumstances (the conditions a) are something which has actually happened (or will unavoidably happen) and which cannot be changed afterwards. Ideal worlds in which $\neg a$ is true are therefore excluded. However, some worlds may still be better than others, and there should then be an obligation to make 'the best out of the sad circumstances". Consequently, Hansson presents a possible worlds semantics in which all worlds are ordered by a preference (betterness) relation. $O(x|a)$ is then defined true if x is true in the best a-worlds. Here, we intend to employ semantics that do not make use of any prohairetic betterness relation, but that model deontic operators with regard to given sets of norms and facts.

Challenge 9. *How to define dyadic deontic operators with regard to given sets of norms and facts?*

Input/output logic assumes a set of (conditional) norms G, and a set of unalterable facts A. The facts A may describe a situation that is inconsistent with the output $out(G, A)$: suppose there is a primary norm $(\top, a) \in G$ and a secondary norm $(\neg a, x) \in G$, i.e. $G = \{(\top, a), (\neg a, x)\}$, and $A = \{\neg a\}$. Though $a \in out(G, A)$, it makes no sense to describe a as obligatory since a cannot be realized any more

in the given situation—no crying over spilt milk. Rather, the output should include only the consequent of the secondary obligation x—it is the best we can make out of these circumstances. To do so, we return to the definitions of $\mathit{maxfamily}(G, A, A)$ as the set of all maximal subsets $H \subseteq G$ such that $\mathit{out}(H, A) \cup A$ is consistent, and the set $\mathit{out}^{\cap}(G, A)$ as the intersection of all outputs from $H \in \mathit{maxfamily}(G, A, A)$, i.e. $\mathit{out}^{\cap}(G, A) = \bigcap\{\mathit{out}(H, A) \mid H \in \mathit{maxfamily}(G, A, A)\}$. We may then define:

$$G \models O(x|a) \quad \text{iff} \quad x \in \mathit{out}^{\cap}(G, \{a\})$$

Thus, relative to the set of norms G, $O(x|a)$ is defined true if x is in the output under a of all maximal sets H of norms such that their output under $\{a\}$ is consistent with a. In the example where $G = \{(\top, a), (\neg a, x)\}$ we therefore obtain $O(x|\neg a)$ but not $O(a|\neg a)$ as being true, i.e. only the consequent of the secondary obligation is described as obligatory in conditions $\neg a$.

In the above definition, the antecedent a of the dyadic formula $O(x|a)$ makes the inputs explicit: the truth definition does not make use of any facts other than a. This may be unwanted; one might consider an input set A of *given* facts, and employ the antecedent a only to denote an additional, *assumed* fact. Still, the output should contradict neither the given nor the assumed facts, and the output should include also the normative consequences x of a norm (a, x) given the assumed fact a. This may be realized by the following definition:

$$G, A \models O(x|a) \quad \text{iff} \quad x \in \mathit{out}^{\cap}(G, A \cup \{a\})$$

So, relative to a set of norms G and a set of facts A, $O(x, a)$ is defined true if x is in the output under $A \cup \{a\}$ of all maximal sets H of norms such that their output under $A \cup \{a\}$ is consistent with $A \cup \{a\}$.

Hansson's description of dyadic deontic operators as describing defeasible obligations that are subject to change when more specific, namely contrary-to-duty situations emerge, may be the most prominent view, but it is by no means the only one. Earlier authors like von Wright [119, 120] and Anderson [5] have proposed more normal conditionals, which in particular support 'strengthening of the antecedent' SA $O(x|a) \to O(x|a \wedge b)$. From an input/output perspective, such operators can be accommodated by defining

$$G, A \models O(x|a) \quad \text{iff} \quad x \in \mathit{out}(G, A \cup \{a\})$$

It is immediate that for all standard output operations out_n this definition validates SA. The properties of dyadic deontic operators that are, like the above, semantically defined within the framework of input/output logic, have not been studied so far. The theorems they validate will inevitably depend on what output operation is chosen, cf. Hansen [51] for some related conjectures.

10 Permissive norms

In formal deontic logic, permission is studied less frequently than obligation. For a long time, it was naively assumed that it can simply be taken as a dual of obligation, just as possibility is the dual of necessity in modal logic. Permission is then defined as the absence of an obligation to the contrary, and the modal operator P defined by $Px =_{def} \neg O\neg x$. Today's focus on obligations is not only in stark contrast how deontic logic began, for when von Wright [117] started modern deontic logic in 1951, it was the P-operator that he took as primitive, and defined obligation as an absence of a permission to the contrary. Rather, more and more authors have come to realize how subtle and multi-faceted the concept of permission is. Much energy was devoted to solving the problem of 'free choice permission', where one may derive from the statement that one is permitted to have a cup of tea or a cup of coffee that it is permitted to have a cup of tea, and it is permitted to have a cup of coffee, or for short, that $P(x \vee y)$ implies Px and Py (cf. Kamp [66]). Von Wright, in his late work starting with [123], dropped the concept of inter-definability of obligations and permissions altogether by introducing P-norms and O-norms, where one may call something permitted only if it derives from the collective contents of some O-norms and at most one P-norm. This concept of 'strong permission' introduced deontic 'gaps': whereas in standard deontic logic SDL, $O\neg x \vee Px$ is a tautology, meaning that any state of affairs is either forbidden or permitted, von Wright's new theory means that in the absence of explicit P-norms only what is obligatory is permitted, and that nothing is permitted if also O-norms are missing. Perhaps most importantly, Bulygin [24] observed that an authoritative kind of permission must be used in the context of multiple authorities and updating normative systems: if a higher authority permits you to do something, a lower authority can no longer prohibit it. Summing up, the understanding of permission is still in a less satisfactory state than the understanding of obligation and prohibition. Indeed, a whole chapter in the handbook of deontic logic and normative systems is devoted to the various forms of permission [54].

Challenge 10. *How to distinguish various kinds of permissions and relate them to obligations?*

From the viewpoint of input/output logic, one may first try to define a concept of negative permission in the line of the classic approach. Such a definition is the following:

$$G, A \models P^{neg}x \quad \text{iff} \quad \neg x \notin out(G, A)$$

So something is permitted by a code iff its negation is not obligatory according to the code and in the given situation. As innocuous and standard as such a definition

seems, questions arise as to what output operation *out* may be used. Simple-minded output out_1 and basic output out_2 produce counterintuitive results: consider a set of norms G of which one norm $(work, tax)$ demands that if I am employed then I have to pay taxes. For the default situation $A = \{\top\}$ then $P^{neg}(work \wedge \neg tax)$ is true, i.e. it is by default permitted that I am employed and do not pay taxes. Stronger output operations out_3 and out_4 that warrant reusable output exclude this result, but their use in deontic reasoning is questionable due to contrary-to-duty reasoning, as discussed in Section 1.

In contrast to a concept of negative permission, one may also define a concept of 'strong' or 'positive permission'. This requires a set P of explicit permissive norms, just as G is a set of explicit obligations. As a first approximation, one may say that something is positively permitted by a code iff the code explicitly presents it as such. However, this leaves a central logical question unanswered as to how explicitly given permissive and obligating norms may generate permissions that—in some sense—follow from the explicitly given norms. Pursuing von Wright's later approach, we may define:

$$G, P \models P^{stat}(x/a) \quad \text{iff} \quad x \in out(G \cup \{(b, y)\}, a) \text{ for some } (b, y) \in P \cup \{(\top, \top)\}$$

So there is a permission to realize x in conditions a if x is generated under these conditions either by the norms in G alone, or the norms in G together with some explicit permission (b, y) in P. We call this a 'static' version of strong permission. For example, consider a set G consisting of the norm $(work, tax)$, and a set P consisting of the sole license $(18y, vote)$ that permits all adults to take part in political elections. Then all of the following are true: $P^{stat}(tax/work)$, $P^{stat}(vote/18y)$, $P^{stat}(tax/work \wedge male)$ and also $P^{stat}(vote/\neg work \wedge 18y)$ (so even unemployed adults are permitted to vote).

Where negative permission is liberal, in the sense that anything is permitted that does not conflict with one's obligations, the concept of static permission is quite strict, as nothing is permitted that does not explicitly occur in the norms. In between, one may define a concept of 'dynamic permission' that defines something as permitted in some situation a if forbidding it for these conditions would prevent an agent from making use of some explicit (static) permission. The formal definition reads:

$$G, P \models P^{dyn}(x/a) \quad \text{iff} \quad \neg y \in out(G \cup \{(a, \neg x)\}, b) \text{ for some } y \text{ and conditions } b \text{ such that } G, P \models P^{stat}(y/b)$$

Consider the above static permission $P^{stat}(vote/\neg work \wedge 18y)$ that even the unemployed adult populations is permitted to vote, generated by $P = \{(18y, vote)\}$ and $G = \{(work, tax)\}$. We might also like to say, without reference to age, that the

unemployed are protected from being forbidden to vote, and in this sense are permitted to vote, but $P^{stat}(vote/\neg work)$ is not true. And we might like to say that adults are protected from being forbidden to vote unless they are employed, and in this sense are permitted to be both unemployed and take part in elections, but also $P^{stat}(\neg work \wedge vote/18y)$ is not true. Dynamic permissions allow us to express such protections, and make both $P^{dyn}(vote/\neg work)$ and $P^{dyn}(\neg work \wedge vote/18y)$ true: if either $(\neg work, \neg vote)$ or $(18y, (\neg work \rightarrow \neg vote))$ were added to G we would obtain $\neg vote$ as output in conditions $(\neg work \wedge 18y)$ in spite of the fact that, as we have seen, $G, P \models P^{stat}(vote/\neg work \wedge 18y)$.

The relation of permission and obligation can also be studied from a multi-agent perspective. Think of two brothers who are fighting for a toy, and the mother obliges the son who's playing with the toy to permit his brother to play as well.

There are, ultimately, a number of questions for all these concepts of permissions that Makinson and van der Torre have further explored [87]. Other kinds of permissions have been discussed from an input/output perspective in the literature, too, for example permissions as exceptions of obligations [13]. It seems input/output logic is able to help clarify the underlying concepts of permission better than traditional deontic semantics. One challenge is Governatori's paradox [39], containing a conditional norm whose body and head are permissions: "the collection of medical information is permitted provided that the collection of personal information is permitted."

11 Meaning postulates and intermediate concepts

To define a deontic operator of individual obligation seems straightforward if the norm in question is an individual command or act of promising. For example, if you are the addressee α of the following imperative sentence

(1) You, hand me that screwdriver, please.

and you consider the command valid, then what you ought to do is to hand the screwdriver in question to the person β uttering the request. In terms of input/output logic, let x be the proposition that α hands the screwdriver to β: with the set of norms $G = \{(\top, x)\}$, the set of facts $A = \{\top\}$, and the truth definition Ox iff $x \in out(A, G)$: then we obtain that Ox is true, i.e. it is true that it ought to be that α hands the screwdriver to β.

Norms that belong to a legal system are more complex, and thus more difficult to reason about. Consider, for example

(2) An act of theft is punished by a prison sentence not exceeding 5 years or a fine.

Things are again easy if you are a judge and you know that the accused in front of you has committed an act of theft—then you ought to hand out a verdict that commits the accused to pay a fine or to serve a prison sentence not exceeding 5 years. However, how does the judge arrive at the conclusion that an act of theft has been committed? 'Theft' is a legal term that is usually accompanied by a legal definition such as the following one:

(3) Someone commits an act of theft if that person has taken a movable object from the possession of another person into his own possession with the intention to own it, and if the act occurred without the consent of the other person or some other legal authorization.

It is noteworthy that (3) is not a norm in the strict sense—it does not prescribe or allow a behavior—but rather a stipulative definition, or, in more general terms, a *meaning postulate* that constitutes the legal meaning of theft. Such sentences are often part of the legal code. They share with norms the property of being neither true nor false: stipulative definitions are neither empirical statements nor descriptive statements. In this sense we say that they are neither true nor false. However, they are held to be true by definition. The significance of (3) is that it decomposes the complex legal term 'theft' into more basic legal concepts. These concepts are again the subject of further meaning postulates, among which may be the following:

(4) A person in the sense of the law is a human being that has been born.

(5) A movable object is any physical object that is not a person or a piece of land.

(6) A movable object is in the possession of a person if that person is able to control the uses and the location of the object.

(7) The owner of an object is—within the limits of the law—entitled to do with it whatever he wants, namely keep it, use it, transfer possession or ownership of the object to another person, and destroy or abandon it.

Not all of definitions (4)-(7) may be found in the legal statutes, though they may be viewed as belonging to the normative system by virtue of having been accepted in legal theory and judicial reasoning. They constitute 'intermediate concepts': they link legal terms (person, movable object, possession etc.) to words describing natural facts (human being, born, piece of land, keep an object etc.).

Any proper representation of legal norms must include means of representing meaning postulates that define legal terms, decompose legal terms into more basic legal terms, or serve as intermediate concepts that link legal terms to terms that describe natural facts. But for deontic logic, with its standard possible worlds semantics, a comprehensive solution to the problem of representing meaning postulates is so far lacking (cf. Lindahl [78]).

Challenge 11. *How can meaning postulates and intermediate terms be modeled in semantics for deontic logic reasoning?*

The representation of intermediate concepts is of particular interest, since such concepts arguably reduce the number of implications required for the transition from natural facts to legal consequences and thus serve an economy of expression (cf. Lindahl and Odelstad [79] and their recent overview chapter [80]). Lindahl and Odelstad use the term 'ownership' as an example to argue as follows: let $F_1, ..., F_p$ be descriptions of some situations in which a person α acquires ownership of an object γ, e.g. by acquiring it from some other person β, finding it, building it from owned materials, etc., and let $C_1, ..., C_n$ be among the legal consequences of α's ownership of γ, e.g. freedom to use the object, rights to compensation when the object is damaged, obligations to maintain the object or pay taxes for it etc. To express that each fact F_i has the consequence C_j, $p \times n$ implications are required. The introduction of the term $\mathit{Ownership}(x, y)$ reduces the number of required implications to $p + n$: there are p implications that link the facts $F_1, ..., F_p$ to the legal term $\mathit{Ownership}(x, y)$, and n implications that link the legal term $\mathit{Ownership}(x, y)$ to each of the legal consequences $C_1, ..., C_n$. The argument obviously does not apply to all cases: one implication $(F_1 \vee ... \vee F_p) \rightarrow (C_1 \wedge ... \wedge C_n)$ may often be sufficient to represent the case that a variety of facts $F_1, ..., F_p$ has the same multitude of legal consequences $C_1, ..., C_n$. However, things may be different when norms that link a number of factual descriptions to the same legal consequences stem from different normative sources, may come into conflict with other norms, can be overridden by norms of higher priority, or be subject to individual exemption by norms that grant freedoms or licenses: in these cases, the norms must be represented individually. So it seems worthwhile to consider ways to incorporate intermediate concepts into a formal semantics for deontic logic.

In an input/output framework, a first step could be to employ a separate set T of theoretical terms, namely meaning postulates, alongside the set G of norms. Let T consists of intermediates of the form (a, x), where a is a factual sentence (e.g. that β is in possession of γ, and that α and β agreed that α should have γ, and that β hands γ to α), and x states that some legal term obtains (e.g. that α is now owner of γ). To derive outputs from the set of norms G, one may then use $A \cup \mathit{out}(T, A)$ as input, i.e. the factual descriptions together with the legal statements that obtain given the intermediates T and the facts A.

It may be of particular interest to see that such a set of intermediates may help resolve possible conflicts in the law. Let $(\top, \neg \mathit{dog})$ be a statute that forbids dogs on the premises, but let there also be a higher order principle that no blind person may be required to give up his or her guide dog. Of course the conflict may be solved

by modifying the statute (e.g. add a condition that the dog in question is not a guide dog), but then modifying a statute is usually not something a judge, faced with such a norm, is allowed to do: the judge's duty is solely to consider the statute, interpret it according to the known or supposed will of the norm-giver, and apply it to the given facts. The judge may then come to the conclusion that a fair and considerate norm-giver would not have meant the statute to apply to guide dogs, i.e. the term "dog" in the statute is a theoretical term whose extension is smaller than the natural term. So the statute must be re-interpreted as reading $(\top, \neg tdog)$ with the additional intermediate $(dog \wedge \neg guidedog, tdog) \in T$, and thus no conflict arises for the case of blind persons that want to keep their guide dog. While this seems to be a rather natural view of how judicial conflict resolution works (the example is taken from an actual court case), the exact process of creating and modifying theoretical terms in order to resolve conflicts must be left to further study.

12 Constitutive norms

Constitutive norms like counts-as conditionals are rules that create the possibility of or define an activity. For example, according to Searle [102], the activity of playing chess is constituted by action in accordance with these rules. Chess has no existence apart from these rules. The institutions of marriage, money, and promising are like the institutions of baseball and chess in that they are systems of such constitutive rules or conventions. They have been identified as the key mechanism to normative reasoning in dynamic and uncertain environments, for example to realize agent communication, electronic contracting, dynamics of organizations, see, e.g., Boella and van der Torre [14].

Challenge 12. *How to define counts-as conditionals and relate them to obligations and permissions?*

For Jones and Sergot [64], the counts-as relation expresses the fact that a state of affairs or an action of an agent "is a sufficient condition to guarantee that the institution creates some (usually normative) state of affairs". They formalize this introducing a conditional connective $\Rightarrow_s$ to express the "counts-as" connection in the context of an institution s. They characterize the logic of $\Rightarrow_s$ as a conditional logic, with axioms for agglomeration $((x \Rightarrow_s y) \,\&\, (x \Rightarrow_s z)) \supset (x \Rightarrow_s (y \wedge z))$, left disjunction $((x \Rightarrow_s z) \,\&\, (y \Rightarrow_s z)) \supset ((x \vee y) \Rightarrow_s z)$ together with transitivity $((x \Rightarrow_s y) \,\&\, (y \Rightarrow_s z)) \supset (x \Rightarrow_s z)$. The flat fragment can be phrased as an input/output logic as follows [15].

Definition 12.1. *Let L be a propositional action logic with $\vdash$ the related notion of derivability and Cn the related consequence operation $Cn(x) = \{y \mid x \vdash y\}$. Let CA be a set of pairs of L, $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, read as '$x_1$ counts as y_1', etc. Moreover, consider the following proof rules conjunction for the output (AND), disjunction of the input (OR), and transitivity (T) defined as follows:*

$$\frac{(x, y_1), (x, y_2)}{(x, y_1 \wedge y_2)} AND \qquad \frac{(x_1, y), (x_2, y)}{(x_1 \vee x_2, y)} OR \qquad \frac{(x, y_1), (y_1, y_2)}{(x, y_2)} T$$

For an institution s, the counts-as output operator out_{CA} is defined as the closure operator on the set CA using the rules above together with a tacit rule that allows replacement of logical equivalents in input and output. We write $(x, y) \in out_{\mathrm{CA}}(CA, s)$. Moreover, for $X \subseteq L$, we write $y \in out_{\mathrm{CA}}(CA, s, X)$ if there is a finite $X' \subseteq X$ such that $(\wedge X', y) \in out_{\mathrm{CA}}(CA, s)$, indicating that the output y is derived by the output operator for the input X, given the counts-as conditionals CA of institution s. We also write $out_{\mathrm{CA}}(CA, s, x)$ for $out_{\mathrm{CA}}(CA, s, \{x\})$.

Example 12.2. *If for some institution s we have $CA = \{(a, x), (x, y)\}$, then we have $out_{\mathrm{CA}}(CA, s, a) = \{x, y\}$.*

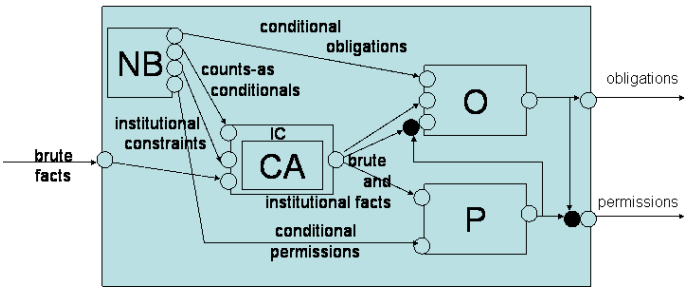
The recognition that statements like "X counts as Y in context c" may have different meanings in different situations lead Grossi *et al.* [45, 46] to propose a family of operators capturing four notions of counts-as conditionals. Starting from a simple modal logic of contexts, several logics are used to define the family of operators. All logics have been proven to be sound and strongly complete. By using a logic of acceptance, Lorini *et al.* [81, 82] investigate another aspect of constitutive norms, that is, the fact that agents of a society need to accept such norms in order for them to be in force.

Considering the legal practice, Governatori and Rotolo [40] propose a study of constitutive norms within the framework of defeasible logic. This allows them to capture de defeasibility of counts-as conditionals: even in presence of a constitutive norms like "X counts as Y in context c", the inference of Y from X can be blocked in presence of exceptions.

There is presently no consensus on the logic of counts-as conditionals, probably due to the fact that the concept is not studied in depth yet. For example, the adoption of the transitivity rule T for their logic is criticized by Artosi *et al.* [8]. Jones and Sergot say that "we have been unable to produce any counter-instances [of transitivity], and we are inclined to accept it". Neither of these authors considers replacing transitivity by cumulative transitivity (CT): $((x \Rightarrow_s y) \& (x \wedge y \Rightarrow_s z)) \supset (x \Rightarrow_s z)$, that characterizes operations out_3, out_4 of input/output logic. For a more

comprehensive overview on constitutive norms, the reader is referred to the chapter by Grossi and Jones [44] in the handbook of deontic logic and normative systems.

The main issue in defining constitutive norms like counts-as conditionals is defining their relation to regulative norms like obligations and permissions. Boella and van der Torre [15] use the notion of a logical architecture combining several logics into a more complex logical system, also called logical input/output nets (or *lions*).



The notion of logical architecture naturally extends the input/output logic framework, since each input/output logic can be seen as the description of a 'black box'. In the above figure there are boxes for counts-as conditionals (CA), institutional constraints (IC), obligating norms (O) and explicit permissions (P). The norm base (NB) component contains sets of norms or rules, which are used in the other components to generate the component's output from its input. The figure shows that the counts-as conditionals are combined with the obligations and permissions using iteration, that is, the counts-as conditionals produce institutional facts, which are input for the norms. Roughly, if we write $out(CA, G, A)$ for the output of counts-as conditionals together with obligations, $out(G, A)$ for obligations as before, then $out(CA, G, A) = out(G, out_{CA}(CA, A))$.

There are many open issues concerning constitutive norms, since their logical analysis has not attracted much attention yet. How to distinguish among various kinds of constitutive norms? How are constitutive norms (x counts as y) distinguished from classifications (x is a y)? What is the relation with intermediate concepts?

13 Revision of a set of norms

In general, a code G of regulations is not static, but changes over time. For example, a legislative body may want to introduce new norms or to eliminate some existing ones. A different (but related) type of change is the one induced by the fusion of two (or more) codes—a topic addressed in the next section. A related but different

issue not addressed here is that of how norms come about, how they propagate in the society, and how they change over time.

Little work exists on the logic of the revision of a set of norms. To the best of our knowledge, Alchourrón and Makinson [3, 4] were the first to study the changes of a legal code. The addition of a new norm n causes an enlargement of the code, consisting of the new norm plus all the regulations that can be derived from n. Alchourrón and Makinson distinguish two other types of change. When the new norm is incoherent with the existing ones, we have an *amendment* of the code: in order to coherently add the new regulation, we need to reject those norms that conflict with n. Finally, *derogation* is the elimination of a norm n together with whatever part of G implies n.

Alchourrón and Makinson [3] assume a "hierarchy of regulations". Alchourrón and Bulygin [2] also considered the *Normenordnung* and the consequences of gaps in this ordering. For example, in jurisprudence the existence of precedents is an established method to determine the ordering among norms.

However, although Alchourrón and Makinson aim at defining change operators for a set of norms of some legal system, the only condition they impose on G is that it is a non-empty and finite set of propositions. In other words, a norm x is taken to be simply a formula in propositional logic. Thus, they suggest that "the same concepts and techniques may be taken up in other areas, wherever problems akin to inconsistency and derogation arise" ([3], p. 147).

This explains how their work (together with Gärdenfors's analysis of counterfactuals) could ground that research area that is now known as *belief revision*. Belief revision is the formal study of how a set of propositions changes in view of new information that may be inconsistent with the existing beliefs. Expansion, revision and contraction are the three belief change operations that Alchourrón, Gärdenfors and Makinson identified in their approach (called AGM) and that have a clear correspondence with the changes on a system of norms we mentioned above.

Challenge 13. *How to revise a set of regulations or obligations?*

Recently, AGM theory has been reconsidered as a framework for norm change. However, beside syntactic approaches where norm change is performed directly on the set of norms (as in AGM), there are also proposals that appeared in the dynamic logic literature and that could be described as semantic approaches.

One example of this is the dynamic context logic proposed by Aucher et al. [9], where norm change is a form of model update. Point of depart is a dynamic variant of the logic of context used to study counts-as conditionals introduced by Grossi et al. [46]. Context expansion and context contraction operators are defined. Context expansion and context contraction represent the promulgation and the derogation

of constitutive norms respectively. One of the advantages of this approach is that it can be used for the formal specification and verification of computational models of interactions based on norms.

A formal account clearly rooted in the legal practice is the one proposed by Governatori and Rotolo [41]. In particular, the removal of norms can be performed by annulment or by abrogation. The crucial difference between these two mechanisms is that annulment removes a norm from the code and all its effects (past and future) are cancelled. Abrogation, on the other hand, does not operate retroactively, and so it leaves the effects of an abrogated norm holding in the past.

It should then be clear that, in order to capture the difference between annulment and abrogation, the temporal dimension is pivotal. For this reason, Governatori and Rotolo's first attempt is to use theory revision in Defeasible Logic without temporal reasoning is unsuccessful as it cannot capture retroactivity. They the add a temporal dimension to Defeasible Logic to keep track of the changes in a normative system and to deal with retroactivity. Norms are represented along two temporal dimensions: the time of validity when the norm enters in the normative system and the time of effectiveness when the norm can produce legal effects. This leads to keep multiple versions of a normative system are needed. If Governatori and Rotolo [41] manage to capture the temporal dimension that plays a role in legal modifications, the resulting formalisation is rather complex.

To overcome such complexity without losing hold on the legal practice, Governatori et al. [42] explored three AGM-like contraction operators to remove rules, add exceptions and revise rule priorities.

Boella et al. [12] also use AGM theory, where propositional formulas are replaced by pairs of propositional formulas to represent rules, and the classical consequence operator Cn is replaced by an input/output logic. Within this framework, AGM contraction and revision of rules are studied. It is shown that results from belief base dynamics can be transferred to rule base dynamics. However, difficulties arise in the transfer of AGM theory change to rule change. In particular, it is shown that the six basic postulates of AGM contraction are consistent only for some input/output logics but not for others. Furthermore, it is shown how AGM rule revision can be defined in terms of AGM rule contraction using the Levi identity.

When we turn to a proper representation of norms, as in the input/output logic framework, the AGM principles thus prove to be too general to deal with the revision of a normative system. For example, one difference between revising a set of beliefs and revising a set of regulations is the following: when a new norm is added, coherence may be restored by modifying some of the existing norms, not necessarily retracting some of them. The following example clarifies this point:

Example. If we have $\{(\top, a), (a, b)\}$ and we have that c is an exception to the obligation to do b, then we need to retract (c, b). Two possible solutions are $\{(\neg c, a), (a, b)\}$ or $\{(\top, a), (a \wedge \neg c, b)\}$.

Stolpe [106] also combines input/output logic and AGM theory to propose an abstract model of norm change. Contraction is used to represent the derogation of a norm, that is, the elimination of a norm together with whatever part of the code that implies that norm. This is rendered as an AGM partial meet contraction with a selection function for a set of norms in input/output logic. Stolpe gives a complete AGM-style characterisation of the derogation operation. Revision, on the other hand, serves to study the amendment of a code, which happens when we wish to add a new norm which is incoherent with the existing ones. Amendment is defined as a norm revision obtained via the Levi identity.

Future research must investigate whether general patterns in the revision and contraction of norms exist and how to formalize them. Another open question is whether other logics can offer a general framework for modelling norm change. Finally, more case studies showing that formally defined operators serve for a conceptual analysis of normative change are needed.

14 Merging sets of norms

We now turn to another type of change, that is the aggregation of regulations. This problem has been only recently addressed in the literature and therefore the findings are still incomplete.

The first noticeable thing is the lack of general agreement about where the norms that are to be aggregated come from:

1. some papers focus on the merging of conflicting norms that belong to the same normative system [29];

2. other papers assume that the regulations to be fused belong to different systems [18]; and finally

3. some authors provide patterns of possible rules to be combined, and consider both cases 1. and 2. above [43].

The first situation seems to be more a matter of coherence of the whole system rather than a genuine problem of fusion of norms. However, such approaches have the merit to reveal the tight connections between fusion of norms, non-monotonic

logics and defeasible deontic reasoning. The initial motivation for the study of belief revision was the ambition to model the revision of a set of regulations. In contrast to this, the generalization of belief revision to *belief merging* is primarily dictated by the goal to tackle the problem—arising in computer science—of combining information from different sources. The pieces of information are represented in a formal language and the aim is to merge them in an (ideally) unique knowledge base. See Konieczny and Grégoire [71] for a survey on logic-based approaches to information fusion.

Challenge 14. *Can the belief merging framework deal with the problem of merging sets of norms?*

If, following Alchourrón and Makinson, we assume that norms are unconditional, then we could expect to use standard merging operators to fuse sets of norms. Yet once we consider conditional norms, as in the input/output logic framework, problems arise again. Moreover, most of the fusion procedures proposed in the literature seem to be inadequate for the scope.

To see why this is the case, we need to explain the merging approach in a few words. Let us assume that we have a finite number of belief bases $K_1, K_2, \ldots, K_n$ to merge. IC is the belief base whose elements are the integrity constraints (i.e., any condition that we want the final outcome to satisfy). Given a multi-set $E = \{K_1, K_2, \ldots, K_n\}$ and IC, a merging operator $\mathcal{F}$ is a function that assigns a belief base to E and IC. Let $\mathcal{F}_{IC}(E)$ be the resulting collective base from the IC fusion on E.

Fusion operators come in two types: model-based and syntax-based. The idea of a model-based fusion operator is that models of $\mathcal{F}_{IC}(E)$ are models of IC, which are preferred according to some criterion depending on E. Usually the preference information takes the form of a total pre-order on the interpretations induced by a notion of distance $d(w, E)$ between an interpretation w and E.

Syntax-based merging operators are usually based on the selection of some consistent subsets of $\bigcup E$ [10, 70]. The bases K_i in E can be inconsistent and the result does not depend on the distribution of the well formed formulas over the members of the group. Konieczny [70] refers the term 'combination' to the syntax-based fusion operators to distinguish them from the model-based approaches.

Finally, the model-based aggregation operators for bases of equally reliable sources can be of two sorts. On the one hand, there are majoritarian operators that are based on a principle of distance-minimization [77]. On the other hand, there are egalitarian operators, which look at the distribution of the distances in E [69]. These two types of merging try to capture two intuitions that often guide the aggregation of individual preferences into a social one. One option is to let the majority decide the collective outcome, and the other possibility is to equally distribute

the individual dissatisfaction.

Obviously, these intuitions may well serve in the aggregation of individual knowledge bases or individual preferences, but have nothing to say when we try to model the fusion of sets of norms. Hence, for this purpose, syntactic merging operators may be more appealing. Nevertheless, the selection of a coherent subset depends on additional information like an order of priority over the norms to be merged, or some other meta-principles.

The reader may wonder about the relationships between merging sets of norms and the revision of a normative system. In particular, one may speculate that Challenge 14 is not independent of Challenge 13, and that a positive answer to Challenge 14 implies an answer to 13. This is indeed an interesting question, but we believe that the answer to this question is not straightforward. Konieczny and Pino Pérez [72] have shown that there are close links between belief merging operators and belief revision ones. In particular, they show that an IC merging operator is an extension of an AGM revision operator. However, as we have seen, it is not clear whether IC merging operators could be properly used to study the merging of norms.

An alternative approach is to generalize existing belief change operators to merging rules. This is the approach followed by Booth et al. [18], where merging operators defined using a consolidation operation and possibilistic logic are applied to the aggregation of conditional norms in an input/output logic framework. However, at this preliminary stage, it is not clear whether such methodology is more fruitful for testing the flexibility of existing operators to tackle other problems than the ones they were created for, or if this approach can really shed some light on the new riddle at hand.

Grégoire [43] takes a different perspective. Here, real examples from the Belgian-French bilateral agreement preventing double taxation are considered. These are fitted into a taxonomy of the most common legal rules with exceptions, and the combination of each pair of norms is analyzed. Moreover, both the situations in which the regulations come from the same system and those in which they come from different ones are contemplated, and some general principles are derived. Finally, a merging operator for rules with abnormality propositions is proposed. A limitation of Grégoire's proposal is that only the aggregation of rules with the same consequence is taken into account and, in our opinion, this neglects other sorts of conflicts that may arise, as we see now.

Cholvy and Cuppens [29] also call for non-monotonic reasoning in the treatment of contradictions, and present a method for merging norms. The proposal assumes an order of priority among the norms to be merged and this order is used to resolve the incoherence. Even though this is quite a strong assumption, Cholvy and Cuppens's

work takes into consideration a broader type of incoherence than Grégoire [43]. In their example, an organization that works with secret documents has two rules. R_1 is "It is obligatory that any document containing some secret information is kept in a safe, when nobody is using this document". R_2 is "If nobody has used a given document for five years, then it is obligatory to destroy this document by burning it". As they observe, in order to deduce that the two rules are conflicting, we need to introduce the constraint that keeping a document and destroying it are contradictory actions. That is, the notion of coherence between norms can involve information not given by any norms.

15 Games, norms and obligations

Deontic logic has been developed as a logic for practical reasoning, and normative systems are used to guide, control, or regulate desired system behaviour. This raises a number of questions. For example, how are deontic logic and the logic of normative systems related to alternative decision and agent interaction models such as BDI theory, decision theory, game theory, or social choice theory? Moreover, how can deontic logic be extended with cognitive concepts such as beliefs, desires, goals, intentions, and commitments? Though there have been a few efforts to base deontic logic with a logic of knowledge to define knowledge based obligation [92], or to extend deontic logic with BDI concepts [20], we believe that such extensions have not been fully explored yet. For example, Kolodny and MacFarlane [68] describe a decision problem involving miners, as well as several dialogues scenarios, which highlight the problems of normative reasoning with agents.

Maybe the most fundamental challenge has become apparent in this article. We discussed how deontic STIT logics are based on interactions of agents in games, and we discussed how norm based deontic logics have been developed on the basis of detachment. However, these two approaches have not been combined yet. So this is our final challenge in this article.

Challenge 15. *How can deontic logic be based on both norm and detachment, as well as decision and game theory?*

Norms and games have been related before. Lewis [76] introduced master-slave games and Bulygin [24] introduced Rex-minister-subject games in a discussion on the role of permissive norms in normative systems and deontic logic. Moreover, deontic logic has been used as an element in games to partially influence the behavior of individual agents [17]. Van der Torre [109] proposes games as the foundation of deontic logic. He illustrate the notion of a violation game using a metaphor from

daily life. A person faces the parental problem of letting the son go to bed in time, or letting him make his homework. The mother is obliging her son to eat his vegetables. As illustrated in the first drawing of Figure 5, the son did what his mother asked him to do.

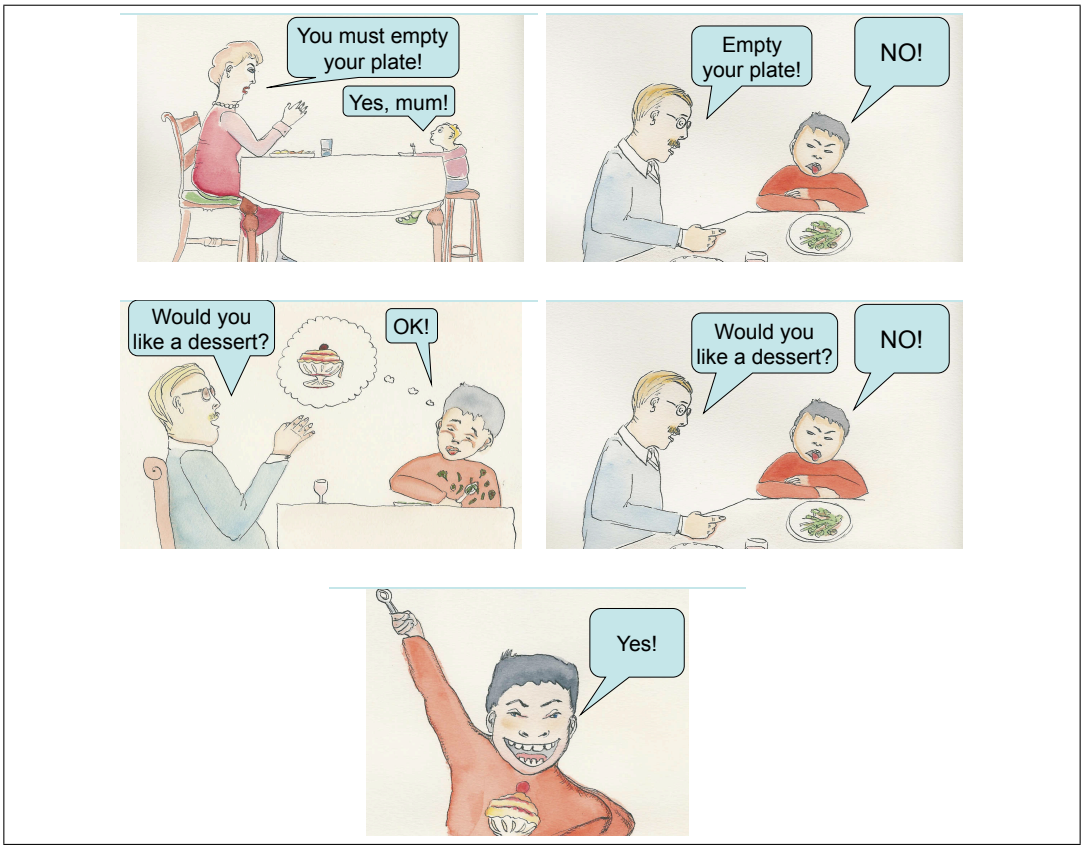


Figure 5: Conformance, violation, incentive, violation, negotiation (Drawings by Egberdien van der Torre), from [van der Torre, 2010].

However, in the second drawing his behavior has changed. The son does not like vegetables, and when the parents tell the boy to eat his vegetables, he just says "No!" At the third drawing, when the son's desire not to eat vegetables became stronger than his motivation to obey his parents, the parents adapted their strategy and introduced the use of incentives. They told their son, "if you empty your plate you will get a dessert", or sometimes, "if you don't finish your plate, you don't get a dessert." The boy has a desire to eat a dessert, and this desire is stronger than the desire not to eat vegetables, so he is eating his vegetables again. However, after

some time we reach the fourth figure where the incentive no longer works. The boy starts to protest and to negotiate. In those cases, the parents sometimes decide that the son will get his dessert even without eating his vegetables, for example, because the child still has eaten at least some of them, or because it is his birthday, or simply because they are not in the mood to argue. As visualized in the fifth figure, this makes the boy very happy. It is precisely this aspect that characterizes a violation game. The violation does not follow necessarily from the norm, but is subject to exceptions and negotiation.

Figure 6 models this example by a standard extensive game tree. Let's look first at one moment in time. The child decides first whether to eat his vegetables or not. But in this decision, he takes the response of his parents into account. In other words, he has a model of how the parents will respond to his behavior. In the deontic logic we propose here, based on a violation game, it is obligatory to empty the plate when the boy expects that not eating his vegetables leads to violation, not when a violation logically follows. By the way, we identify the recognition of violation and the sanction in the example for illustrative purposes, in reality usually two distinct steps can be distinguished.

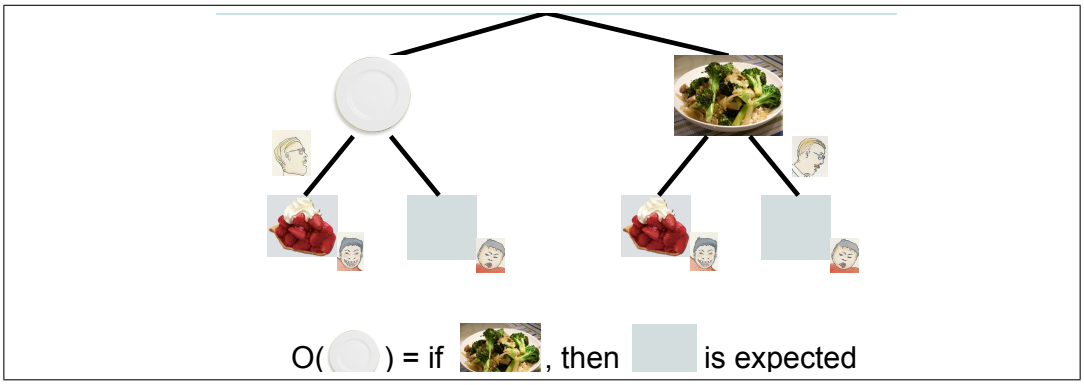


Figure 6: Expectation, from [van der Torre, 2010]

The general definition of obligation based on violation games extends this basic idea to behavior over periods of time. Let's consider the three phases in the example. Borrowing from terminology from classical game theory, we say that it is obligatory to eat the vegetables, when not eating them and the strategy that this leads to a violation, is an equilibrium. In the first phase in which the son eats his vegetables, the violation is only implicit since it does not occur. In the second phase not eating the vegetables is identified with the absence of the dessert. In the third phase, the boy may sometimes eat his vegetables, and sometimes not. As long as the norm is in force, he will still believe to be sanctioned most of the time when he does not eat

his vegetables. When the sanction is not applied most of the time we have reached a fourth phase, in which we say that the norm is no longer in force.

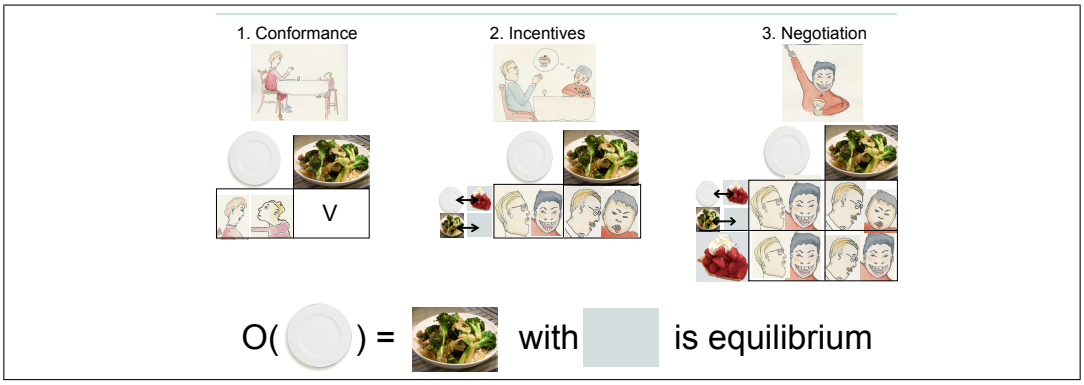


Figure 7: Equilibrium, from [van der Torre, 2010]

Summarizing, norms are rules defining a violation game.

Definition 15.1 (Violation games [109])**.** *Violation games are social interactions among agents to determine whether violations have occurred, and which sanctions will be imposed for such violations. A normative system is a specification of violation games.*

Since norms do not have truth values, we cannot say that two normative systems are logically equivalent, or that a normative system implies a norm. Therefore it has been proposed to take equivalence of normative systems as the fundamental principle of deontic logic. Implication is then replaced by acceptance and redundancy, which are defined in terms of norm equivalence: a norm is accepted by a normative system if adding it to the normative system leads to an equivalent normative system, and a norm is redundant in a normative system if removing it from the normative system leads to an equivalent normative system. The fundamental notion of equivalence of normative systems can be defined in terms of violation games.

Definition 15.2 (Equivalence of normative systems [109])**.** *Two normative systems are equivalent if and only if they define the same set of violation games.*

Finally, we can now give a more precise definition of an autonomous system. Remember that auto means self, and nomos means norm.

Definition 15.3 (Autonomy [109])**.** *A system is autonomous if and only if it can play violation games.*

Violation games are the basis of normative reasoning and deontic logic, but more complex games must be considered too. Consider for example the following situation. If a child is in the water and there is one bystander, chances are that the bystander will jump into the water and save the child. However, if there are one hundred bystanders, chances are that no-one jumps in the water and the child will drown. How to reason about such bystander effects?

Van der Torre suggests that an extension of violation games, called norm creation games [17], may be used to analyze the situation. An agent reasons as follows. What is the explicit norm I would like to adopt for such situations? Clearly, if I would be in the water and I could not swim, or it is my child drowning in the water, then I would like prefer that someone would jump in the water. To be precise, I would accept a norm that in such cases, the norm for each individual would be to jump into the water. Consequently, one should act according to this norm, and everyone should jump into the water. Norm creation games can be used to give a more general definition of a normative system.

Definition 15.4 (Norm creation games [109])**.** *Norm creation games are social interactions among agents to determine which norms are in force, whether norm violations have occurred, and which sanctions will be imposed for such violations. A normative system is a specification of norm creation games.*

There are many details to be further discussed here. For example, if there is a way to discriminate among the people and it may be assumed that all people would follow this discrimination, then only some people have to jump into the water (the men, the good swimmers if they can be identified, the tall people, and so on). In general, and as common in legal reasoning, the more that is known about the situation, the more can be said about the protocol leading to the norm.

For the semantics of the new deontic logic founded on violation games, one needs a way to derive obligations from norms, as in the iterative detachment approach, or input/output logic. The extension now is to represent the agents and their games into the semantic structures, and derive the norms from that using game theoretic methods. As the norm creation game illustrates, also protocols for norm creation must be represented to model more complex games.

The language of the new deontic logic founded on violation games will be richer than most of the deontic logics studied thus far. There will be formal statements referring to the regulative, permissive and constitutive norms, as in the input/output logic framework, but there will also be an explicit representation of the games the agents are playing. Many choices are possible here, and the area of game theory will lead the way.

We need other approaches that represent norms and obligations at the same time, since deontic logic founded on violation games has to built on it. We also also have to study time, actions, mental modalities, permissions and constitutive norms, since they all play a role in violation games. We also need a precise understanding of Anderson's idea of violation conditions which do not necessarily lead to sanctions, but to the more abstract notion of "a bad state," i.e. a state in which something bad has happened. Whereas many of these deontic problems have been studied in isolation in the deontic logic literature, we believe that violation games will work as a metaphor to bring these problems together, and study their interdependencies.

16 Summary

The aim of this article is to introduce readers to the area of deontic logic and its challenges. The interested reader is advised to download the handbook of deontic logic and normative systems, and should not take our article only as its guidance. In particular, in this article we have not gone into the formal aspects of deontic logic. Deontic logicians have developed monadic modal logics, non-monotonic ones, rule based systems, and much more. The formalisms developed in deontic logic have also been adopted by a wider logic community, in particular the preference based deontic logics have been adopted in many areas [83].

As far as open problems are concerned, in the context of the handbook this concerns mainly the problems of *multiagent* deontic logic and problems related to normative systems. We have addressed the following challenges.

How to reconstruct the history of traditional deontic logic as a challenge to deal with contrary to duty reasoning, violations and preference (Challenge 1)?

What are the challenges in game theoretic approach to normative reasoning (Section 2), which is based on non-deterministic actions (Challenge 2), moral luck (Challenge 3) and procrastination (Challenge 4)?

How to reconstruct the history of modern deontic logic as a challenge to deal with Jørgensen's dilemma and detachment (Challenge 5), and more generally to bridge the tradition of normative system with the tradition of modal deontic logic?

What is the challenge in multi agent detachment of obligations from norms? For example, when detaching obligations from norms, when do agents assume that other agents comply with their norms (Challenge 6)? In game theory, agents assume that other agents are rational in the sense of acting in their best interest. Analogously, multiagent deontic logic raises the question when agents assume that other agents comply with their norms. For answering the question, we assume that every norm is directed towards a single agent, and that the normative system does not change.

How do norm based semantics handle the traditional challenges in deontic logic? These problems are when a set of norms may be termed 'coherent' (Challenge 7), how to deal with normative conflicts (Challenge 8), how to interpret dyadic deontic operators that formalize 'it ought to be that x on conditions α' as $O(x/\alpha)$ (Challenge 9), how various concepts of permission can be accommodated (Challenge 10), how meaning postulates and counts-as conditionals can be taken into account (Challenge 11 and 12), and how sets of norms may be revised and merged (Challenge 13 and 14).

Finally, how can the two approaches of game based deontic logic and norm based deontic logic be combined? (Challenge 15)

References

[1] Thomas Ågotnes, Wiebe van der Hoek, and Michael Wooldridge. Robust normative systems and a logic of norm compliance. *Logic Journal of the IGPL*, 18(1):4–30, 2010.

[2] C. E. Alchourrón and E. Bulygin. The expressive conception of norms. In R. Hilpinen, editor, *New Studies in Deontic Logic*, pages pp 95–124. Reidel, Dordrecht, 1981.

[3] C. E. Alchourrón and D. Makinson. Hierarchies of regulations and their logic. In R. Hilpinen, editor, *New Studies in Deontic Logic*, pages 125–148. Reidel, Dordrecht, 1981.

[4] C. E. Alchourrón and D. Makinson. On the logic of theory change: Contraction functions and their associated revision functions. *Theoria*, 48:14–37, 1982.

[5] A. R. Anderson. On the logic of commitment. *Philosophical Studies*, 19:23–27, 1959.

[6] G. Andrighetto, G. Governatori, P. Noriega, and L. van der Torre, editors. *Normative Multi-Agent Systems*, volume 4 of *Dagstuhl Follow-Ups*. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2013.

[7] L. Åqvist. Good Samaritans, contrary-to-duty imperatives and epistemic obligations. *Noûs*, 1:361–379, 1967.

[8] A. Artosi, A. Rotolo, and S. Vida. On the logical nature of count-as conditionals. In *The Law of Electronic Agents. Proceedings of the LEA04 Workshop*, pages 9–33, Bologna, 2004.

[9] G. Aucher, D. Grossi, A. Herzig, and E. Lorini. Dynamic context logic. In X. He, J. Horty, and E. Pacuit, editors, *Logic, Rationality, and Interaction: Second International Workshop, LORI 2009, Chongqing, China, October 8-11, 2009. Proceedings*, pages 15–26, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[10] C. Baral, S. Kraus, J. Minker, and V. S. Subrahmanian. Combining knowledge bases consisting of first-order theories. *Computational Intelligence*, 8:45–71, 1992.

[11] G. Boella, J. M. Broersen, and L. van der Torre. Reasoning about constitutive norms, counts-as conditionals, institutions, deadlines and violations. In The Duy Bui, Tuong Vinh Ho, and Quang-Thuy Ha, editors, *Intelligent Agents and Multi-Agent*

Systems, 11th Pacific Rim International Conference on Multi-Agents, PRIMA 2008, Hanoi, Vietnam, December 15-16, 2008. Proceedings, volume 5357 of *Lecture Notes in Computer Science*, pages 86–97. Springer, 2008.

[12] G. Boella, G. Pigozzi, and L. van der Torre. AGM contraction and revision of rules. *Journal of Logic, Language and Information*, 25(3-4):273–297, 2016.

[13] G. Boella and L. van der Torre. Permissions and obligations in hierarchical normative systems. In *Proceedings of the 9th International Conference on Artificial Intelligence and Law, ICAIL 2003, Edinburgh, Scotland, UK, June 24-28, 2003*, pages 109–118, Edinburgh, 2003.

[14] G. Boella and L. van der Torre. Constitutive norms in the design of normative multiagent systems. In *Computational Logic in Multi-Agent Systems, 6th International Workshop, CLIMA VI*, LNCS 3900, pages 303–319. Springer, 2006.

[15] G. Boella and L. van der Torre. A logical architecture of a normative system. In *Deontic Logic and Artificial Normative Systems, 8th International Workshop on Deontic Logic in Computer Science (DEON'06)*, volume 4048 of *LNCS*, pages 24–35, Berlin, 2006. Springer.

[16] G. Boella, L. van der Torre, and H. Verhagen. Introduction to normative multiagent systems. *Computation and Mathematical Organizational Theory, special issue on normative multiagent systems*, 12(2-3):71–79, 2006.

[17] Guido Boella and Leendert van der Torre. A game-theoretic approach to normative multi-agent systems. In *Normative Multi-agent Systems, 18.03. - 23.03.2007*, Dagstuhl Seminar Proceedings, 2007.

[18] R. Booth, S. Kaci, and L. van der Torre. Merging rules: Preliminary version. In *Proceedings of the Eleventh International Workshop on Non-Monotonic Reasoning (NMR'06)*, 2006.

[19] C. Boutilier. Toward a logic for qualitative decision theory. In *Proceedings of the 4th International Conference on Principles of Knowledge Representation and Reasoning (KR'94). Bonn, Germany, May 24-27, 1994.*, pages 75–86, Bonn, 1994.

[20] J. Broersen, M. Dastani, and L. van der Torre. BDIOCTL: Obligations and the specification of agent behavior. In Georg Gottlob and Toby Walsh, editors, *IJCAI-03, Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, Acapulco, Mexico, August 9-15, 2003*, pages 1389–1390. Morgan Kaufmann, 2003.

[21] J. Broersen and L. van der Torre. Reasoning about norms, obligations, time and agents. In A. K. Ghose, G. Governatori, and R. Sadananda, editors, *Agent Computing and Multi-Agent Systems, 10th Pacific Rim International Conference on Multi-Agents, PRIMA 2007, Bangkok, Thailand, November 21-23, 2007. Revised Papers*, volume 5044 of *Lecture Notes in Computer Science*, pages 171–182. Springer, 2007.

[22] J. M. Broersen, F. Dignum, V. Dignum, and J-J Ch. Meyer. Designing a deontic logic of deadlines. In A. Lomuscio and D. Nute, editors, *Deontic Logic in Computer Science, 7th International Workshop on Deontic Logic in Computer Science, DEON 2004, Madeira, Portugal, May 26-28, 2004. Proceedings*, volume 3065 of *Lecture Notes in Computer Science*, pages 43–56. Springer, 2004.

[23] J. M. Broersen and L. van der Torre. What an agent ought to do. *Artif. Intell. Law*, 11(1):45–61, 2003.

[24] E. Bulygin. Permissive norms and normative systems. In A. Martino and F. Socci Natali, editors, *Automated Analysis of Legal Texts*, pages 211–218. Publishing Company, Amsterdam, 1986.

[25] J. Carmo and A.J.I. Jones. Deontic logic and contrary-to-duties. In D.Gabbay and F.Guenthner, editors, *Handbook of Philosophical Logic*, volume 8, pages 265–343. Kluwer, 2002.

[26] H-N Castañeda. The paradoxes of deontic logic: The simplest solution to all of them in one fell swoop. In R. Hilpinen, editor, *New Studies in Deontic Logic: Norms, Actions, and the Foundations of Ethics*, pages 37–85. Springer Netherlands, Dordrecht, 1981.

[27] H-N Castañeda. The logical structure of legal systems: A new perspective. In A. A. Martino, editor, *Deontic Logic, Computational Linguistics and Legal Information Systems, volume II*, page 21?37. North Holland Publishing, Dordrecht, 1982.

[28] R.M. Chisholm. Contrary-to-duty imperatives and deontic logic. *Analysis*, 24:33Ű36, 1963.

[29] L. Cholvy and F. Cuppens. Reasoning about norms provided by conflicting regulations. In P. McNamara and H. Prakken, editors, *Norms, Logics and Information Systems*, pages 247–264. IOS, Amsterdam, 1999.

[30] L. Cholvy and C. Garion. An attempt to adapt a logic of conditional preferences for reasoning with contrary-to-duties. *Fundam. Inform.*, 48(2-3):183–204, 2001.

[31] C. Condoravdi and S. Lauer. Anankastic conditionals are just conditionals. *Semantics and Pragmatics*, 9(8):1–69, November 2016.

[32] S. Danielsson. *Preference and Obligation: Studies in the Logic of Ethics*. Filosofiska fÃűreningen, Uppsala, 1968.

[33] J. Forrester. Gentle murder, or the adverbial Samaritan. *Journal of Philosophy*, 81:193–196, 1984.

[34] D. Gabbay, J. Horty, X. Parent, R. van der Meyden, and L. van der Torre, editors. *Handbook of Deontic Logic and Normative Systems*. College Publications, London, UK, 2013.

[35] L. Goble. Multiplex semantics for deontic logic. *Nordic Journal of Philosophical Logic*, 5:113–134, 2000.

[36] L. Goble. A logic for deontic dilemmas. *Journal of Applied Logic*, 3:461–483, 2005.

[37] L. Goble. Normative conflicts and the logic of 'ought'. *Noûs*, 43:450–489, 2009.

[38] L. Goble. Prima facie norms, normative conflicts, and dilemmas. In D. Gabbay, J. Horty, X. Parent, R. van der Meyden, and L. van der Torre, editors, *Handbook of Deontic Logic*, pages 249–352. College Publications, 2013.

[39] G. Governatori. Thou shalt is not you will. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law, ICAIL 2015, San Diego, CA, USA, June 8-12, 2015*, pages 63–68, 2015.

[40] G. Governatori and A. Rotolo. A computational framework for institutional agency.

Artificial Intelligence and Law, 16(1):25–52, 2008.

[41] G. Governatori and A. Rotolo. Changing legal systems: legal abrogations and annulments in defeasible logic. *Logic Journal of IGPL*, 18(1):157–194, 2010.

[42] G. Governatori, A. Rotolo, F. Olivieri, and S. Scannapieco. Legal contractions: A logical analysis. In E. Francesconi and B. Verheij, editors, *Proceedings of the 9th International Conference on Artificial Intelligence and Law, ICAIL 2003, Edinburgh, Scotland, UK, June 24-28, 2003*, pages 63–72. ACM, 2013.

[43] E. Grégoire. Fusing legal knowledge. In *Proceedings of the 2004 IEEE Int. Conf. on Information Reuse and Integration (IEEE-IRI'2004)*, pages 522–529, 2004.

[44] D. Grossi and A. Jones. Constitutive norms and counts-as conditionals. In D. Gabbay, J. Horty, X. Parent, R. van der Meyden, and L. van der Torre, editors, *Handbook of Deontic Logic and Normative Systems*, pages 407–441. College Publications, 2013.

[45] D. Grossi, J-J Ch. Meyer, and F. Dignum. Classificatory aspects of counts-as: An analysis in modal logic. *Journal of Logic and Computation*, 16(5):613–643, 2006.

[46] D. Grossi, J-J Ch. Meyer, and F. Dignum. The many faces of counts-as: A formal analysis of constitutive-rules. *Journal of Applied Logic*, 6(2):192–217, 2008.

[47] J. Hansen. Sets, sentences, and some logics about imperatives. *Fundamenta Informaticae*, 48:205–226, 2001.

[48] J. Hansen. Problems and results for logics about imperatives. *Journal of Applied Logic*, 2:39–61, 2004.

[49] J. Hansen. Conflicting imperatives and dyadic deontic logic. *Journal of Applied Logic*, 3:484–511, 2005.

[50] J. Hansen. Deontic logics for prioritized imperatives. *Artificial Intelligence and Law*, 3(3-4):484–511, 2005.

[51] J. Hansen. Prioritized conditional imperatives: problems and a new proposal. *Autonomous Agents and Multi-Agent Systems*, 17(1):11–35, 2008.

[52] J. Hansen, G. Pigozzi, and L. van der Torre. Ten philosophical problems in deontic logic. In *Normative Multi-agent Systems, 18.03. - 23.03.2007*, Dagstuhl Seminar Proceedings, 2007.

[53] B. Hansson. An analysis of some deontic logics. *Nôus*, 3:373–398, 1969. Reprinted in [56] 121–147.

[54] S. O. Hansson. The varieties of permission. In D. Gabbay, J. Horty, X. Parent, R. van der Meyden, and L. van der Torre, editors, *Handbook of Deontic Logic and Normative Systems*, pages 195–240. College Publications, 2013.

[55] R. M. Hare. *Moral Thinking*. Clarendon Press, Oxford, 1981.

[56] R. Hilpinen, editor. *Deontic Logic: Introductory and Systematic Readings*. Reidel, Dordrecht, 1971.

[57] R. Hilpinen and P. McNamara. Deontic logic: A historical survey and introduction. In D. Gabbay, J. Horty, X. Parent, R. van der Meyden, and L. van der Torre, editors, *Handbook of Deontic Logic and Normative Systems*, pages 3–136. College Publications, 2013.

[58] J. Horty. *Agency and deontic logic*. Oxford University Press, 2001.

[59] J. F. Horty. Moral dilemmas and nonmonotonic logic. *Journal of Philosophical Logic*, 23:35–65, 1994.

[60] J. F. Horty. Nonmonotonic foundations for deontic logic. In D. Nute, editor, *Defeasible Deontic Logic*, pages 17–44. Kluwer, Dordrecht, 1997.

[61] J. F. Horty. Reasoning with moral conflicts. *Noûs*, 37:557–605, 2003.

[62] J. F. Horty. *Reasons and Defaults*. Oxford University Press, 2012.

[63] F. Jackson and R. Pargetter. Oughts, options, and actualism. *Philosophical Review*, 99:233–255, 1986.

[64] A. Jones and M. Sergot. A formal characterisation of institutionalised power. *Journal of IGPL*, 3:427–443, 1996.

[65] J. Jørgensen. Imperatives and logic. *Erkenntnis*, 7:288–296, 1938.

[66] H. Kamp. Free choice permission. *Proceedings of the Aristotelian Society*, 74:57–74, 1973.

[67] S. Kanger. New foundations for ethical theory: Part 1. duplic., 42 p., 1957. Reprinted in [56] 36–58.

[68] N. Kolodny and J. MacFarlane. Ifs and oughts. *Journal of Philosophy*, 107(3):115–143, 2010.

[69] S. Konieczny. *Sur la Logique du Changement: Révision et Fusion de Bases de Connaissance*. PhD thesis, University of Lille, France, 1999.

[70] S. Konieczny. On the difference between merging knowledge bases and combining them. In *KR 2000, Principles of Knowledge Representation and Reasoning Proceedings of the Seventh International Conference, Breckenridge, Colorado, USA, April 11-15, 2000.*, volume 8, pages 135–144. Morgan Kaufmann, 2000.

[71] S. Konieczny and E. Grégoire. Logic-based approaches to information fusion. *Information Fusion*, 7:4–18, 2006.

[72] S. Konieczny and Ramón P. Pérez. Logic based merging. *Journal of Philosophical Logic*, 40(2):239–270, 2011.

[73] E. J. Lemmon. Moral dilemmas. *The Philosophical Review*, 70:139–158, 1962.

[74] D. Lewis. *Counterfactuals*. Basil Blackwell, Oxford, 1973.

[75] D. Lewis. Semantic analyses for dyadic deontic logic. In S. Stenlund, editor, *Logical Theory and Semantic Analysis*, pages 1 – 14. Reidel, Dordrecht, 1974.

[76] D. Lewis. A problem with permission. In E. Saarinen, R. Hilpinen, I. Niiniluoto, and M. P. Hintikka, editors, *Essays in Honour of Jaako Hintikka: On the Occasion of His Fiftieth Birthday on January 12, 1979*, pages 163–175. Reidel, Dordrecht, 1979.

[77] J. Lin and A. Mendelzon. Merging databases under constraints. *International Journal of Cooperative Information Systems*, 7:55–76, 1996.

[78] L. Lindahl. Norms, meaning postulates, and legal predicates. In E. Garzón Valdés, editor, *Normative Systems in Legal and Moral Theory. Festschrift for Carlos E. Alchourrón and Eugenio Bulygin*, pages 293–307. Duncker & Humblot, Berlin, 1997.

[79] L. Lindahl and J. Odelstad. Intermediate concepts in normative systems. In L. Goble and J-J Ch. Meyer, editors, *Deontic Logic and Artificial Normative Systems: 8th International Workshop on Deontic Logic in Computer Science, DEON 2006, Utrecht, The Netherlands, July 12-14, 2006. Proceedings*, pages 187–200. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[80] L. Lindahl and J. Odelstad. The theory of joining-systems. In D. Gabbay, J. Horty, X. Parent, R. van der Meyden, and L. van der Torre, editors, *Handbook of Deontic Logic and Normative Systems*, pages 545–634. College Publications, 2013.

[81] E. Lorini and D. Longin. A logical account of institutions: From acceptances to norms via legislators. In *Proceedings of the Eleventh International Conference on Principles of Knowledge Representation and Reasoning*, KR'08, pages 38–48. AAAI Press, 2008.

[82] E. Lorini, D. Longin, B. Gaudou, and A. Herzig. The logic of acceptance: Grounding institutions on agents' attitudes. *Journal of Logic and Computation*, 19(6):901–940, 2009.

[83] D. Makinson. Five faces of minimality. *Studia Logica*, 52:339–379, 1993.

[84] D. Makinson. On a fundamental problem of deontic logic. In P. McNamara and H. Prakken, editors, *Norms, Logics and Information Systems. New Studies on Deontic Logic and Computer Science*, pages 29–54. IOS Press, 1999.

[85] D. Makinson and L. van der Torre. Input-output logics. *Journal of Philosophical Logic*, 29(4):383–408, 2000.

[86] D. Makinson and L. van der Torre. Constraints for input-output logics. *Journal of Philosophical Logic*, 30(2):155–185, 2001.

[87] D. Makinson and L. van der Torre. Permissions from an input-output perspective. *Journal of Philosophical Logic*, 32(4):391–416, 2003.

[88] D. Makinson and L. van der Torre. What is input/output logic? In B. Löwe, W. Malzkom, and T. Räsch, editors, *Foundations of the Formal Sciences II : Applications of Mathematical Logic in Philosophy and Linguistics (Papers of a conference held in Bonn, November 10-13, 2000)*, Trends in Logic, vol. 17, pages 163–174, Dordrecht, 2003. Kluwer. Reprinted in this volume.

[89] R. B. Marcus. Moral dilemmas and consistency. *Journal of Philosophy*, 77:121?136, 1980.

[90] I. Niiniluoto. Hypothetical imperatives and conditional obligation. *Synthese*, 66:111–133, 1986.

[91] Loes Olde Loohuis. Obligations in a responsible world. In H. Xiangdong, J. F. Horty, and E. Pacuit, editors, *LORI*, volume 5834 of *Lecture Notes in Computer Science*, pages 251–262. Springer, 2009.

[92] E. Pacuit, R. Parikh, and E. Cogan. The logic of knowledge based obligation. *Synthese*, 149(2):311–341, 2006.

[93] X. Parent. On the strong completeness of åqvist's dyadic deontic logic G. In R. van der Meyden and L. van der Torre, editors, *Deontic Logic in Computer Science, 9th International Conference, DEON 2008, Luxembourg, Luxembourg, July 15-18, 2008. Pro-*

ceedings, volume 5076 of *Lecture Notes in Computer Science*, pages 189–202. Springer, 2008.

[94] X. Parent and L. van der Torre. Aggregative deontic detachment for normative reasoning. In C. Baral, G. De Giacomo, and T. Eiter, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Fourteenth International Conference, KR 2014, Vienna, Austria, July 20-24, 2014*. AAAI Press, 2014.

[95] X. Parent and L. van der Torre. "sing and dance!" - input/output logics without weakening. In *Deontic Logic and Normative Systems - 12th International Conference, DEON 2014, Ghent, Belgium, July 12-15, 2014. Proceedings*, pages 149–165, 2014.

[96] X. Parent and L van der Torre. The pragmatic oddity in norm-based deontic logics. In *Proceedings of The 16th International Conference on Artificial Intelligence and Law*, 2017.

[97] J. Pearl. From conditional oughts to qualitative decision theory. In D. Heckerman and E. H. Mamdani, editors, *UAI '93: Proceedings of the Ninth Annual Conference on Uncertainty in Artificial Intelligence, The Catholic University of America, Providence, Washington, DC, USA, July 9-11, 1993*, pages 12–22. Morgan Kaufmann, 1993.

[98] H. Prakken and M. Sergot. Contrary-to-duty obligations. *Studia Logica*, 57:91–115, 1996.

[99] A. Ross. Imperatives and logic. *Theoria*, 7:53–71, 1941. Reprinted in *Philosophy of Science* **11**:30–46, 1944.

[100] W. D. Ross. *The Right and the Good*. Clarendon Press, Oxford, 1930.

[101] J.-P. Sartre. *L'Existentialisme est un Humanisme*. Nagel, Paris, 1946.

[102] J.R. Searle. *The Construction of Social Reality*. The Free Press, New York, 1995.

[103] T. J. Smiley. The logical basis of ethics. *Acta Philosophica Fennica*, 16:237–246, 1963.

[104] W. Spohn. An analysis of Hansson's dyadic deontic logic. *Journal of Philosophical Logic*, 4:237–252, 1975.

[105] E. Stenius. The principles of a logic of normative systems. *Acta Philosophica Fennica*, 16:247–260, 1963.

[106] A. Stolpe. Norm-system revision: Theory and application. *Artificial Intelligence and Law*, 18:247–283, 2010.

[107] X. Sun and L. van der Torre. Combining constitutive and regulative norms in input/output logic. In *Deontic Logic and Normative Systems - 12th International Conference, DEON 2014, Ghent, Belgium, July 12-15, 2014. Proceedings*, pages 241–257, 2014.

[108] J. van Benthem, D. Grossi, and F. Liu. Priority structures in deontic logic. *Theoria*, 80(2):116–152, 2014.

[109] L. van der Torre. Violation games: a new foundation for deontic logic. *Journal of Applied Non-Classical Logics*, 20(4):457–477, 2010.

[110] L. van der Torre and Y. Tan. The temporal analysis of chisholm's paradox. In J. Mostow and C. Rich, editors, *Proceedings of the Fifteenth National Conference on Artificial Intelligence and Tenth Innovative Applications of Artificial Intelligence*

Conference, AAAI 98, IAAI 98, July 26-30, 1998, Madison, Wisconsin, USA., pages 650–655. AAAI Press / The MIT Press, 1998.

[111] L. van der Torre and Y. Tan. An update semantics for prima facie obligations. In *Proceedings of The 17th European Conference on Artificial Intelligence*, pages 38–42, 1998.

[112] L. van der Torre and Y. Tan. Rights, duties and commitments between agents. In T. Dean, editor, *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, IJCAI 99, Stockholm, Sweden, July 31 - August 6, 1999. 2 Volumes, 1450 pages*, pages 1239–1246. Morgan Kaufmann, 1999.

[113] L. van der Torre and Y. Tan. An update semantics for defeasible obligations. In K. B. Laskey and H. Prade, editors, *UAI '99: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, Stockholm, Sweden, July 30 - August 1, 1999*, pages 631–638. Morgan Kaufmann, 1999.

[114] B. van Fraassen. Values and the heart's command. *Journal of Philosophy*, 70:5–19, 1973.

[115] B. C. van Fraassen. The logic of conditional obligation. *Journal of Philosophical Logic*, 1:417–438, 1972.

[116] G. H. von Wright. Deontic logic. *Mind*, 60:1–15, 1951.

[117] G. H. von Wright. *An Essay in Modal Logic*. North-Holland, Amsterdam, 1951.

[118] G. H. von Wright. A note on deontic logic and derived obligation. *Mind*, 65:507–509, 1956.

[119] G. H. von Wright. A new system of deontic logic. *Danish Yearbook of Philosophy*, 1:173–182, 1961. Reprinted in [56] 105–115.

[120] G. H. von Wright. A correction to a new system of deontic logic. *Danish Yearbook of Philosophy*, 2:103–107, 1962. Reprinted in [56] 115–119.

[121] G. H. von Wright. *Norm and Action*. Routledge & Kegan Paul, London, 1963.

[122] G. H. von Wright. *An Essay in Deontic Logic and the General Theory of Action*. North Holland, Amsterdam, 1968.

[123] G. H. von Wright. Norms, truth and logic. In G. H. von Wright, editor, *Practical Reason: Philosophical Papers vol. I*, pages 130–209. Blackwell, Oxford, 1983.

[124] G.H von Wright. *Logical Studies*. Routledge and Kegan, London, 1957.

[125] Z. Ziemba. Deontic syllogistics. *Studia Logica*, 28:139–159, 1971.

 Received 30 October 2016